This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2001 (27.12.2001)

PCT

(10) International Publication Number
WO 01/98318 A1

(51) International Patent Classification[7]: C07H 21/02, 21/04, C12Q 1/68, C12N 5/00, 5/02, 15/00, 15/09, 15/63, 15/70, 15/74, 15/85, 15/87, A01N 43/04, A61K 31/70

(21) International Application Number: PCT/US01/20028

(22) International Filing Date: 22 June 2001 (22.06.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/213,284 22 June 2000 (22.06.2000) US

(71) Applicant *(for all designated States except US)*: **GENAISSANCE PHARMACEUTICALS, INC.** [US/US]; Five Science Park, New Haven, CT 06511 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **KLIEM, Stefanie, E.** [DE/US]; 1298 Hartford Turnpike, Apt. 4E, North Haven, CT 06473 (US). **KOSHY, Beena** [IN/US]; 1298 Hartford Turnpike, Apt. 11B, North Haven, CT 06473 (US). **TANGUAY, Debra, A.** [US/US]; 902 Aspen Glen Drive, Hamden, CT 06518 (US).

(74) Agents: **SHTIVELBAND, Inna** et al.; Genaissance Pharmaceuticals, Inc., Five Science Park, New Haven, CT 06511 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*
— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/98318 A1

(54) Title: HAPLOTYPES OF THE SLC26A2 GENE

(57) Abstract: Novel genetic variants of the Solute Carrier Family 26, Member 2 (SLC26A2) gene are described. Various genotypes, haplotypes, and haplotype pairs that exist in the general United States population are disclosed for the SLC26A2 gene. Compositions and methods for haplotyping and/or genotyping the SLC26A2 gene in an individual are also disclosed. Polynucleotides defined by the haplotypes disclosed herein are also described.

HAPLOTYPES OF THE SLC26A2 GENE

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Serial No. 60/213,284 filed June 22, 2000.

FIELD OF THE INVENTION

This invention relates to variation in genes that encode pharmaceutically-important proteins. In particular, this invention provides genetic variants of the human solute carrier family 26, member 2 (SLC26A2) gene and methods for identifying which variant(s) of this gene is/are possessed by an individual.

BACKGROUND OF THE INVENTION

Current methods for identifying pharmaceuticals to treat disease often start by identifying, cloning, and expressing an important target protein related to the disease. A determination of whether an agonist or antagonist is needed to produce an effect that may benefit a patient with the disease is then made. Then, vast numbers of compounds are screened against the target protein to find new potential drugs. The desired outcome of this process is a lead compound that is specific for the target, thereby reducing the incidence of the undesired side effects usually caused by activity at non-intended targets. The lead compound identified in this screening process then undergoes further *in vitro* and *in vivo* testing to determine its absorption, disposition, metabolism and toxicological profiles. Typically, this testing involves use of cell lines and animal models with limited, if any, genetic diversity.

What this approach fails to consider, however, is that natural genetic variability exists between individuals in any and every population with respect to pharmaceutically-important proteins, including the protein targets of candidate drugs, the enzymes that metabolize these drugs and the proteins whose activity is modulated by such drug targets. Subtle alteration(s) in the primary nucleotide sequence of a gene encoding a pharmaceutically-important protein may be manifested as significant variation in expression, structure and/or function of the protein. Such alterations may explain the relatively high degree of uncertainty inherent in the treatment of individuals with a drug whose design is based upon a single representative example of the target or enzyme(s) involved in metabolizing the drug. For example, it is well-established that some drugs frequently have lower efficacy in some individuals than others, which means such individuals and their physicians must weigh the possible benefit of a larger dosage against a greater risk of side effects. Also, there is significant variation in how well people metabolize drugs and other exogenous chemicals, resulting in substantial interindividual variation in the toxicity and/or efficacy of such exogenous substances (Evans et al., 1999, *Science* 286:487-491). This variability in efficacy or toxicity of a drug in genetically-diverse patients makes many drugs ineffective or even dangerous in certain groups of the population, leading to the failure of such drugs

in clinical trials or their early withdrawal from the market even though they could be highly beneficial for other groups in the population. This problem significantly increases the time and cost of drug discovery and development, which is a matter of great public concern.

It is well-recognized by pharmaceutical scientists that considering the impact of the genetic variability of pharmaceutically-important proteins in the early phases of drug discovery and development is likely to reduce the failure rate of candidate and approved drugs (Marshall A 1997 *Nature Biotech* **15**:1249-52; Kleyn PW et al. 1998 *Science* **281**: 1820-21; Kola I 1999 *Curr Opin Biotech* **10**:589-92; Hill AVS et al. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 62-76; Meyer U.A. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 41-49; Kalow W et al. 1999 *Clin. Pharm. Therap.* **66**:445-7; Marshall, E 1999 *Science* **284**:406-7; Judson R et al. 2000 *Pharmacogenomics* **1**:1-12; Roses AD 2000 *Nature* **405**:857-65). However, in practice this has been difficult to do, in large part because of the time and cost required for discovering the amount of genetic variation that exists in the population (Chakravarti A 1998 *Nature Genet* **19**:216-7; Wang DG et al 1998 *Science* **280**:1077-82; Chakravarti A 1999 *Nat Genet* **21**:56-60 (suppl); Stephens JC 1999 *Mol. Diagnosis* **4**:309-317; Kwok PY and Gu S 1999 *Mol. Med. Today* **5**:538-43; Davidson S 2000 *Nature Biotech* **18**:1134-5).

The standard for measuring genetic variation among individuals is the haplotype, which is the ordered combination of polymorphisms in the sequence of each form of a gene that exists in the population. Because haplotypes represent the variation across each form of a gene, they provide a more accurate and reliable measurement of genetic variation than individual polymorphisms. For example, while specific variations in gene sequences have been associated with a particular phenotype such as disease susceptibility (Roses AD *supra*; Ulbrecht M et al. 2000 *Am J Respir Crit Care Med* **161**: 469-74) and drug response (Wolfe CR et al. 2000 *BMJ* **320**:987-90; Dahl BS 1997 *Acta Psychiatr Scand* **96** (Suppl 391): 14-21), in many other cases an individual polymorphism may be found in a variety of genomic backgrounds, i.e., different haplotypes, and therefore shows no definitive coupling between the polymorphism and the causative site for the phenotype (Clark AG et al. 1998 *Am J Hum Genet* **63**:595-612; Ulbrecht M et al. 2000 *supra*; Drysdale et al. 2000 *PNAS* **97**:10483-10488). Thus, there is an unmet need in the pharmaceutical industry for information on what haplotypes exist in the population for pharmaceutically-important genes. Such haplotype information would be useful in improving the efficiency and output of several steps in the drug discovery and development process, including target validation, identifying lead compounds, and early phase clinical trials (Marshall et al., *supra*).

One pharmaceutically-important gene for the treatment of osteochondrodysplasias is the solute carrier family 26, member 2 (SLC26A2) gene or its encoded product. The transport of sulfates into connective tissue cells, especially chondrocytes, is predominantly dependent upon the transporter encoded by the SLC26A2 gene (OMIM entry: 222600). Sulfate transport is an integral factor in the normal formation and maintenance of cartilage and bone, wherein a steady supply of sulfates is

necessary for the synthesis of the chondroitin sulfate chains attached to connective tissue proteoglycans. The resulting matrix creates a viscous gel that is largely responsible for the ability of cartilage and bone to absorb large compressive loads (Watanabe, et al., 1998. *J. Biochem.* (Tokyo), 124:687-93).

Impairment of sulfate transport across the cell membrane leads to insufficient sulfation of cartilage proteoglycans, thereby diminishing the sulfate content of cartilage and disrupting the process of endochondral bone formation (Satoh H, et al. 1998. *J. Biol. Chem.* 273(20):12307-15; Sperti-Furga et al., 1996, *Am. J. Med. Genet.* 63:144-7). A substantial body of evidence exists demonstrating that mutations in SLC26A2, in particular, constitute a pleiotropic family of recessively inherited osteochondrodysplasias including achondrogenesis type 1B, atelosteogenesis type II, and diastrophic dysplasia (Rossi A, et al. 1998. *Matrix Biol.* 17(5):361-9; Satoh H, et al. *Supra*). These osteochondrodysplasias exhibit a range of pathological severity and comprise a diverse spectrum of clinical presentations. Distinguishing features of these disorders include scoliosis, clubbed feet, cleft palate, congenital heart defects and shortened, malformed limbs and digits characteristic of diastrophic dwarfism (OMIM entry: 222600).

The solute carrier family 26, member 2 gene is located on chromosome 5q31-q34 and contains 2 exons that encode a 739 amino acid protein. A reference sequence for the SLC26A2 gene is shown in Figure 1(Genaissance Contig No. 3758668; SEQ ID NO: 1). Reference sequences for the coding sequence (GenBank Accession No. NM_000112.1) and protein are shown in Figures 2 (SEQ ID NO: 2) and 3 (SEQ ID NO: 3), respectively.

There is one single nucleotide polymorphism in SLC26A2 which has been reported previously in the literature (NCBI SNP ID: rs30832). This polymorphism corresponds to the site named PS4 herein, consisting of a cytosine or thymine at nucleotide position 140013 in Figure 1. This variation is expressed in the coding sequence at nucleotide position 1721 in Figure 2, giving rise to either a threonine or isoleucine variant at amino acid position 574 in Figure 3.

Because of the potential for variation in the SLC26A2 gene to affect the expression and function of the encoded protein, it would be useful to know whether additional polymorphisms exist in the SLC26A2 gene, as well as how such polymorphisms are combined in different copies of the gene. Such information could be applied for studying the biological function of SLC26A2 as well as in identifying drugs targeting this protein for the treatment of disorders related to its abnormal expression or function.

SUMMARY OF THE INVENTION

Accordingly, the inventors herein have discovered 4 novel polymorphic sites in the SLC26A2 gene. These polymorphic sites (PS) correspond to the following nucleotide positions in Figure 1: 136098 (PS1), 136195 (PS2), 139338 (PS3) and 140357 (PS5). The polymorphisms at these sites are guanine or adenine at PS1, adenine or guanine at PS2, thymine or adenine at PS3 and adenine or

thymine at PS5. In addition, the inventors have determined the identity of the alleles at these sites, as well as at the previously identified site at nucleotide position 140013 (PS4), in a human reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: African descent, Asian, Caucasian and Hispanic/Latino. From this information, the inventors deduced a set of haplotypes and haplotype pairs for PS1-PS5 in the SLC26A2 gene, which are shown below in Tables 5 and 4, respectively. Each of these SLC26A2 haplotypes defines a naturally-occurring isoform (also referred to herein as an "isogene") of the SLC26A2 gene that exists in the human population. The frequency with which each haplotype and haplotype pair occurs within the total reference population and within each of the four major population groups included in the reference population was also determined.

Thus, in one embodiment, the invention provides a method, composition and kit for genotyping the SLC26A2 gene in an individual. The genotyping method comprises identifying the nucleotide pair that is present at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3 and PS5 in both copies of the SLC26A2 gene from the individual. A genotyping composition of the invention comprises an oligonucleotide probe or primer which is designed to specifically hybridize to a target region containing, or adjacent to, one of these novel SLC26A2 polymorphic sites. A genotyping kit of the invention comprises a set of oligonucleotides designed to genotype each of these novel SLC26A2 polymorphic sites. In a preferred embodiment, the genotyping kit comprises a set of oligonucleotides designed to genotype each of PS1-PS5. The genotyping method, composition, and kit are useful in determining whether an individual has one of the haplotypes in Table 5 below or has one of the haplotype pairs in Table 4 below.

The invention also provides a method for haplotyping the SLC26A2 gene in an individual. In one embodiment, the haplotyping method comprises determining, for one copy of the SLC26A2 gene, the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3 and PS5. In another embodiment, the haplotyping method comprises determining whether one copy of the individual's SLC26A2 gene is defined by one of the SLC26A2 haplotypes shown in Table 5, below, or a sub-haplotype thereof. In a preferred embodiment, the haplotyping method comprises determining whether both copies of the individual's SLC26A2 gene are defined by one of the SLC26A2 haplotype pairs shown in Table 4 below, or a sub-haplotype pair thereof. The method for establishing the SLC26A2 haplotype or haplotype pair of an individual is useful for improving the efficiency and reliability of several steps in the discovery and development of drugs for treating diseases associated with SLC26A2 activity, e.g., osteochondrodysplasias.

For example, the haplotyping method can be used by the pharmaceutical research scientist to validate SLC26A2 as a candidate target for treating a specific condition or disease predicted to be associated with SLC26A2 activity. Determining for a particular population the frequency of one or more of the individual SLC26A2 haplotypes or haplotype pairs described herein will facilitate a decision on whether to pursue SLC26A2 as a target for treating the specific disease of interest. In

particular, if variable SLC26A2 activity is associated with the disease, then one or more SLC26A2 haplotypes or haplotype pairs will be found at a higher frequency in disease cohorts than in appropriately genetically matched controls. Conversely, if each of the observed SLC26A2 haplotypes are of similar frequencies in the disease and control groups, then it may be inferred that variable SLC26A2 activity has little, if any, involvement with that disease. In either case, the pharmaceutical research scientist can, without *a priori* knowledge as to the phenotypic effect of any SLC26A2 haplotype or haplotype pair, apply the information derived from detecting SLC26A2 haplotypes in an individual to decide whether modulating SLC26A2 activity would be useful in treating the disease.

The claimed invention is also useful in screening for compounds targeting SLC26A2 to treat a specific condition or disease predicted to be associated with SLC26A2 activity. For example, detecting which of the SLC26A2 haplotypes or haplotype pairs disclosed herein are present in individual members of a population with the specific disease of interest enables the pharmaceutical scientist to screen for a compound(s) that displays the highest desired agonist or antagonist activity for each of the most frequent SLC26A2 isoforms present in the disease population. Thus, without requiring any *a priori* knowledge of the phenotypic effect of any particular SLC26A2 haplotype or haplotype pair, the claimed haplotyping method provides the scientist with a tool to identify lead compounds that are more likely to show efficacy in clinical trials.

The method for haplotyping the SLC26A2 gene in an individual is also useful in the design of clinical trials of candidate drugs for treating a specific condition or disease predicted to be associated with SLC26A2 activity. For example, instead of randomly assigning patients with the disease of interest to the treatment or control group as is typically done now, determining which of the SLC26A2 haplotype(s) disclosed herein are present in individual patients enables the pharmaceutical scientist to distribute SLC26A2 haplotypes and/or haplotype pairs evenly to treatment and control groups, thereby reducing the potential for bias in the results that could be introduced by a larger frequency of a SLC26A2 haplotype or haplotype pair that had a previously unknown association with response to the drug being studied in the trial. Thus, by practicing the claimed invention, the scientist can more confidently rely on the information learned from the trial, without first determining the phenotypic effect of any SLC26A2 haplotype or haplotype pair.

In another embodiment, the invention provides a method for identifying an association between a trait and a SLC26A2 genotype, haplotype, or haplotype pair for one or more of the novel polymorphic sites described herein. The method comprises comparing the frequency of the SLC26A2 genotype, haplotype, or haplotype pair in a population exhibiting the trait with the frequency of the SLC26A2 genotype or haplotype in a reference population. A higher frequency of the SLC26A2 genotype, haplotype, or haplotype pair in the trait population than in the reference population indicates the trait is associated with the SLC26A2 genotype, haplotype, or haplotype pair. In preferred embodiments, the trait is susceptibility to a disease, severity of a disease, the staging of a disease or response to a drug. In a particularly preferred embodiment, the SLC26A2 haplotype is selected from

the haplotypes shown in Table 5, or a sub-haplotype thereof. Such methods have applicability in developing diagnostic tests and therapeutic treatments for osteochondrodysplasias.

In yet another embodiment, the invention provides an isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the SLC26A2 gene or a fragment thereof. The reference sequence comprises the contiguous sequences shown in Figure 1 (SEQ ID NO:1) and the polymorphic variant comprises at least one polymorphism selected from the group consisting of adenine at PS1, guanine at PS2, adenine at PS3 and thymine at PS5. In a preferred embodiment, the polymorphic variant comprises an additional polymorphism of thymine at PS4.

A particularly preferred polymorphic variant is an isogene of the SLC26A2 gene. A SLC26A2 isogene of the invention comprises guanine or adenine at PS1, adenine or guanine at PS2, thymine or adenine at PS3, cytosine or thymine at PS4 and adenine or thymine at PS5. The invention also provides a collection of SLC26A2 isogenes, referred to herein as a SLC26A2 genome anthology.

In another embodiment, the invention provides a polynucleotide comprising a polymorphic variant of a reference sequence for a SLC26A2 cDNA or a fragment thereof. The reference sequence comprises SEQ ID NO:2 (Fig.2) and the polymorphic cDNA comprises at least one polymorphism selected from the group consisting of adenine at a position corresponding to nucleotide 1046 and thymine at a position corresponding to nucleotide 2065. In a preferred embodiment, the polymorphic variant comprises an additional polymorphism of thymine at a position corresponding to nucleotide 1721. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a SLC26A2 isogene defined by haplotypes 1, 2, 4 and 5.

Polynucleotides complementary to these SLC26A2 genomic and cDNA variants are also provided by the invention. It is believed that polymorphic variants of the SLC26A2 gene will be useful in studying the expression and function of SLC26A2, and in expressing SLC26A2 protein for use in screening for candidate drugs to treat diseases related to SLC26A2 activity.

In other embodiments, the invention provides a recombinant expression vector comprising one of the polymorphic genomic variants operably linked to expression regulatory elements as well as a recombinant host cell transformed or transfected with the expression vector. The recombinant vector and host cell may be used to express SLC26A2 for protein structure analysis and drug binding studies.

In yet another embodiment, the invention provides a polypeptide comprising a polymorphic variant of a reference amino acid sequence for the SLC26A2 protein. The reference amino acid sequence comprises SEQ ID NO:3 (Fig.3) and the polymorphic variant comprises at least one variant amino acid selected from the group consisting of tyrosine at a position corresponding to amino acid position 349 and serine at a position corresponding to amino acid position 689. In some embodiments, the polymorphic variant also comprises isoleucine at a position corresponding to amino acid position 574. A polymorphic variant of SLC26A2 is useful in studying the effect of the variation on the biological activity of SLC26A2 as well as on the binding affinity of candidate drugs targeting SLC26A2 for the treatment of osteochondrodysplasias.

The present invention also provides antibodies that recognize and bind to the above polymorphic SLC26A2 protein variant. Such antibodies can be utilized in a variety of diagnostic and prognostic formats and therapeutic methods.

The present invention also provides nonhuman transgenic animals comprising one of the SLC26A2 polymorphic genomic variants described herein and methods for producing such animals. The transgenic animals are useful for studying expression of the SLC26A2 isogenes *in vivo*, for *in vivo* screening and testing of drugs targeted against SLC26A2 protein, and for testing the efficacy of therapeutic agents and compounds for osteochondrodysplasias in a biological system.

The present invention also provides a computer system for storing and displaying polymorphism data determined for the SLC26A2 gene. The computer system comprises a computer processing unit; a display; and a database containing the polymorphism data. The polymorphism data includes the polymorphisms, the genotypes and the haplotypes identified for the SLC26A2 gene in a reference population. In a preferred embodiment, the computer system is capable of producing a display showing SLC26A2 haplotypes organized according to their evolutionary relationships.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a reference sequence for the SLC26A2 gene (Genaissance Reference No. 3758668; contiguous lines; SEQ ID NO:1), with the start and stop positions of each region of coding sequence indicated with a bracket ([or]) and the numerical position below the sequence and the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence. SEQ ID NO:24 is equivalent to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25). SEQ ID NO:25 is a modified version of SEQ ID NO:24 that shows the context sequence of each polymorphic site, PS1-PS5, in a uniform format to facilitate electronic searching. For each polymorphic site, SEQ ID NO:25 contains a block of 60 bases of the nucleotide sequence encompassing the centrally-located polymorphic site at the 30th position, followed by 60 bases of unspecified sequence to represent that each PS is separated by genomic sequence whose composition is defined elsewhere herein.

Figure 2 illustrates a reference sequence for the SLC26A2 coding sequence (contiguous lines; SEQ ID NO:2), with the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence.

Figure 3 illustrates a reference sequence for the SLC26A2 protein (contiguous lines; SEQ ID NO:3), with the variant amino acid(s) caused by the polymorphism(s) of Figure 2 positioned below the polymorphic site in the sequence.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is based on the discovery of novel variants of the SLC26A2 gene. As described in more detail below, the inventors herein discovered 5 isogenes of the SLC26A2 gene by characterizing the SLC26A2 gene found in genomic DNAs isolated from an Index Repository that contains immortalized cell lines from one chimpanzee and 93 human individuals. The human individuals included a reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: Caucasian (21 individuals), African descent (20 individuals), Asian (20 individuals), or Hispanic/Latino (18 individuals). To the extent possible, the members of this reference population were organized into population subgroups by their self-identified ethnogeographic origin as shown in Table 1 below.

Table 1. Population Groups in the Index Repository

Population Group	Population Subgroup	No. of Individuals
African descent		20
	Sierra Leone	1
Asian		20
	Burma	1
	China	3
	Japan	6
	Korea	1
	Philippines	5
	Vietnam	4
Caucasian		21
	British Isles	3
	British Isles/Central	4
	British Isles/Eastern	1
	Central/Eastern	1
	Eastern	3
	Central/Mediterranean	1
	Mediterranean	2
	Scandinavian	2
Hispanic/Latino		18
	Caribbean	8
	Caribbean (Spanish Descent)	2
	Central American (Spanish Descent)	1
	Mexican American	4
	South American (Spanish Descent)	3

In addition, the Index Repository contains three unrelated indigenous American Indians (one from each of North, Central and South America), one three-generation Caucasian family (from the CEPH Utah cohort) and one two-generation African-American family.

The SLC26A2 isogenes present in the human reference population are defined by haplotypes for 5 polymorphic sites in the SLC26A2 gene, 4 of which are believed to be novel. The SLC26A2 polymorphic sites identified by the inventors are referred to as PS1-PS5 to designate the order in which they are located in the gene (see Table 3 below), with the novel polymorphic sites referred to as

PS1, PS2, PS3 and PS5. Using the genotypes identified in the Index Repository for PS1-PS5 and the methodology described in the Examples below, the inventors herein also determined the pair of haplotypes for the SLC26A2 gene present in individual human members of this repository. The human genotypes and haplotypes found in the repository for the SLC26A2 gene include those shown in Tables 4 and 5, respectively. The polymorphism and haplotype data disclosed herein are useful for validating whether SLC26A2 is a suitable target for drugs to treat osteochondrodysplasias, screening for such drugs and reducing bias in clinical trials of such drugs.

In the context of this disclosure, the following terms shall be defined as follows unless otherwise indicated:

Allele - A particular form of a genetic locus, distinguished from other forms by its particular nucleotide sequence.

Candidate Gene – A gene which is hypothesized to be responsible for a disease, condition, or the response to a treatment, or to be correlated with one of these.

Gene - A segment of DNA that contains all the information for the regulated biosynthesis of an RNA product, including promoters, exons, introns, and other untranslated regions that control expression.

Genotype – An unphased 5′ to 3′ sequence of nucleotide pair(s) found at one or more polymorphic sites in a locus on a pair of homologous chromosomes in an individual. As used herein, genotype includes a full-genotype and/or a sub-genotype as described below.

Full-genotype – The unphased 5′ to 3′ sequence of nucleotide pairs found at all polymorphic sites examined herein in a locus on a pair of homologous chromosomes in a single individual.

Sub-genotype – The unphased 5′ to 3′ sequence of nucleotides seen at a subset of the polymorphic sites examined herein in a locus on a pair of homologous chromosomes in a single individual.

Genotyping – A process for determining a genotype of an individual.

Haplotype – A 5′ to 3′ sequence of nucleotides found at one or more polymorphic sites in a locus on a single chromosome from a single individual. As used herein, haplotype includes a full-haplotype and/or a sub-haplotype as described below.

Full-haplotype – The 5′ to 3′ sequence of nucleotides found at all polymorphic sites examined herein in a locus on a single chromosome from a single individual.

Sub-haplotype – The 5′ to 3′ sequence of nucleotides seen at a subset of the polymorphic sites examined herein in a locus on a single chromosome from a single individual.

Haplotype pair – The two haplotypes found for a locus in a single individual.

Haplotyping – A process for determining one or more haplotypes in an individual and includes use of family pedigrees, molecular techniques and/or statistical inference.

Haplotype data - Information concerning one or more of the following for a specific gene: a listing of the haplotype pairs in each individual in a population; a listing of the different haplotypes in

a population; frequency of each haplotype in that or other populations, and any known associations between one or more haplotypes and a trait.

Isoform – A particular form of a gene, mRNA, cDNA or the protein encoded thereby, distinguished from other forms by its particular sequence and/or structure.

Isogene – One of the isoforms of a gene found in a population. An isogene contains all of the polymorphisms present in the particular isoform of the gene.

Isolated – As applied to a biological molecule such as RNA, DNA, oligonucleotide, or protein, isolated means the molecule is substantially free of other biological molecules such as nucleic acids, proteins, lipids, carbohydrates, or other material such as cellular debris and growth media. Generally, the term "isolated" is not intended to refer to a complete absence of such material or to absence of water, buffers, or salts, unless they are present in amounts that substantially interfere with the methods of the present invention.

Locus - A location on a chromosome or DNA molecule corresponding to a gene or a physical or phenotypic feature.

Naturally-occurring – A term used to designate that the object it is applied to, e.g., naturally-occurring polynucleotide or polypeptide, can be isolated from a source in nature and which has not been intentionally modified by man.

Nucleotide pair – The nucleotides found at a polymorphic site on the two copies of a chromosome from an individual.

Phased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, phased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is known.

Polymorphic site (PS) – A position within a locus at which at least two alternative sequences are found in a population, the most frequent of which has a frequency of no more than 99%.

Polymorphic variant – A gene, mRNA, cDNA, polypeptide or peptide whose nucleotide or amino acid sequence varies from a reference sequence due to the presence of a polymorphism in the gene.

Polymorphism – The sequence variation observed in an individual at a polymorphic site. Polymorphisms include nucleotide substitutions, insertions, deletions and microsatellites and may, but need not, result in detectable differences in gene expression or protein function.

Polymorphism data – Information concerning one or more of the following for a specific gene: location of polymorphic sites; sequence variation at those sites; frequency of polymorphisms in one or more populations; the different genotypes and/or haplotypes determined for the gene; frequency of one or more of these genotypes and/or haplotypes in one or more populations; any known association(s) between a trait and a genotype or a haplotype for the gene.

Polymorphism Database – A collection of polymorphism data arranged in a systematic or methodical way and capable of being individually accessed by electronic or other means.

Polynucleotide – A nucleic acid molecule comprised of single-stranded RNA or DNA or comprised of complementary, double-stranded DNA.

Population Group – A group of individuals sharing a common ethnogeographic origin.

Reference Population – A group of subjects or individuals who are predicted to be representative of the genetic variation found in the general population. Typically, the reference population represents the genetic variation in the population at a certainty level of at least 85%, preferably at least 90%, more preferably at least 95% and even more preferably at least 99%.

Single Nucleotide Polymorphism (SNP) – Typically, the specific pair of nucleotides observed at a single polymorphic site. In rare cases, three or four nucleotides may be found.

Subject – A human individual whose genotypes or haplotypes or response to treatment or disease state are to be determined.

Treatment - A stimulus administered internally or externally to a subject.

Unphased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, unphased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is not known.

As discussed above, information on the identity of genotypes and haplotypes for the SLC26A2 gene of any particular individual as well as the frequency of such genotypes and haplotypes in any particular population of individuals is expected to be useful for a variety of drug discovery and development applications. Thus, the invention also provides compositions and methods for detecting the novel SLC26A2 polymorphisms and haplotypes identified herein.

The compositions comprise at least one SLC26A2 genotyping oligonucleotide. In one embodiment, a SLC26A2 genotyping oligonucleotide is a probe or primer capable of hybridizing to a target region that is located close to, or that contains, one of the novel polymorphic sites described herein. As used herein, the term "oligonucleotide" refers to a polynucleotide molecule having less than about 100 nucleotides. A preferred oligonucleotide of the invention is 10 to 35 nucleotides long. More preferably, the oligonucleotide is between 15 and 30, and most preferably, between 20 and 25 nucleotides in length. The exact length of the oligonucleotide will depend on many factors that are routinely considered and practiced by the skilled artisan. The oligonucleotide may be comprised of any phosphorylation state of ribonucleotides, deoxyribonucleotides, and acyclic nucleotide derivatives, and other functionally equivalent derivatives. Alternatively, oligonucleotides may have a phosphate-free backbone, which may be comprised of linkages such as carboxymethyl, acetamidate, carbamate, polyamide (peptide nucleic acid (PNA)) and the like (Varma, R. in Molecular Biology and Biotechnology, A Comprehensive Desk Reference, Ed. R. Meyers, VCH Publishers, Inc. (1995), pages 617-620). Oligonucleotides of the invention may be prepared by chemical synthesis using any suitable methodology known in the art, or may be derived from a biological sample, for example, by restriction digestion. The oligonucleotides may be labeled, according to any technique known in the art, including use of radiolabels, fluorescent labels, enzymatic labels, proteins, haptens, antibodies,

sequence tags and the like.

Genotyping oligonucleotides of the invention must be capable of specifically hybridizing to a target region of a SLC26A2 polynucleotide, i.e., a SLC26A2 isogene. As used herein, specific hybridization means the oligonucleotide forms an anti-parallel double-stranded structure with the target region under certain hybridizing conditions, while failing to form such a structure when incubated with a non-target region or a non-SLC26A2 polynucleotide under the same hybridizing conditions. Preferably, the oligonucleotide specifically hybridizes to the target region under conventional high stringency conditions. The skilled artisan can readily design and test oligonucleotide probes and primers suitable for detecting polymorphisms in the SLC26A2 gene using the polymorphism information provided herein in conjunction with the known sequence information for the SLC26A2 gene and routine techniques.

A nucleic acid molecule such as an oligonucleotide or polynucleotide is said to be a "perfect" or "complete" complement of another nucleic acid molecule if every nucleotide of one of the molecules is complementary to the nucleotide at the corresponding position of the other molecule. A nucleic acid molecule is "substantially complementary" to another molecule if it hybridizes to that molecule with sufficient stability to remain in a duplex form under conventional low-stringency conditions. Conventional hybridization conditions are described, for example, by Sambrook J. et al., in Molecular Cloning, A Laboratory Manual, 2nd Edition, Cold Spring Harbor Press, Cold Spring Harbor, NY (1989) and by Haymes, B.D. et al. in Nucleic Acid Hybridization, A Practical Approach, IRL Press, Washington, D.C. (1985). While perfectly complementary oligonucleotides are preferred for detecting polymorphisms, departures from complete complementarity are contemplated where such departures do not prevent the molecule from specifically hybridizing to the target region. For example, an oligonucleotide primer may have a non-complementary fragment at its 5′ end, with the remainder of the primer being complementary to the target region. Alternatively, non-complementary nucleotides may be interspersed into the oligonucleotide probe or primer as long as the resulting probe or primer is still capable of specifically hybridizing to the target region.

Preferred genotyping oligonucleotides of the invention are allele-specific oligonucleotides. As used herein, the term allele-specific oligonucleotide (ASO) means an oligonucleotide that is able, under sufficiently stringent conditions, to hybridize specifically to one allele of a gene, or other locus, at a target region containing a polymorphic site while not hybridizing to the corresponding region in another allele(s). As understood by the skilled artisan, allele-specificity will depend upon a variety of readily optimized stringency conditions, including salt and formamide concentrations, as well as temperatures for both the hybridization and washing steps. Examples of hybridization and washing conditions typically used for ASO probes are found in Kogan et al., "Genetic Prediction of Hemophilia A" in PCR Protocols, A Guide to Methods and Applications, Academic Press, 1990 and Ruaño et al., 87 *Proc. Natl. Acad. Sci. USA* 6296-6300, 1990. Typically, an ASO will be perfectly complementary to one allele while containing a single mismatch for another allele.

Allele-specific oligonucleotides of the invention include ASO probes and ASO primers. ASO probes which usually provide good discrimination between different alleles are those in which a central position of the oligonucleotide probe aligns with the polymorphic site in the target region (e.g., approximately the 7th or 8th position in a 15mer, the 8th or 9th position in a 16mer, and the 10th or 11th position in a 20mer). An ASO primer of the invention has a 3′ terminal nucleotide, or preferably a 3′ penultimate nucleotide, that is complementary to only one nucleotide of a particular SNP, thereby acting as a primer for polymerase-mediated extension only if the allele containing that nucleotide is present. ASO probes and primers hybridizing to either the coding or noncoding strand are contemplated by the invention.

ASO probes and primers listed below use the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25) at the position of the polymorphic site to represent the two alternative allelic variants observed at that polymorphic site.

A preferred ASO probe for detecting SLC26A2 gene polymorphisms comprises a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

AAGTCCTRTACCCAG (SEQ ID NO:4) and its complement,
TTAAGGARAAGGGAC (SEQ ID NO:5) and its complement,
TCTCATTWTGGAAAA (SEQ ID NO:6) and its complement, and
CAATCCCWCTGTGAG (SEQ ID NO:7) and its complement.

A preferred ASO primer for detecting SLC26A2 gene polymorphisms comprises a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

CTTGGGAAGTCCTRT (SEQ ID NO:8); AACTGGCTGGGTAYA (SEQ ID NO:9);
GCTCAATTAAGGARA (SEQ ID NO:10); TCTTATGTCCCTTYT (SEQ ID NO:11);
TTAGCCTCTCATTWT (SEQ ID NO:12); ATGTAGTTTTCCAWA (SEQ ID NO:13);
TCAGTGCAATCCCWC (SEQ ID NO:14); and GAATCCCTCACAGWG (SEQ ID NO:15).

Other genotyping oligonucleotides of the invention hybridize to a target region located one to several nucleotides downstream of one of the novel polymorphic sites identified herein. Such oligonucleotides are useful in polymerase-mediated primer extension methods for detecting one of the novel polymorphisms described herein and therefore such genotyping oligonucleotides are referred to herein as "primer-extension oligonucleotides". In a preferred embodiment, the 3′-terminus of a primer-extension oligonucleotide is a deoxynucleotide complementary to the nucleotide located immediately adjacent to the polymorphic site.

A particularly preferred oligonucleotide primer for detecting SLC26A2 gene polymorphisms by primer extension terminates in a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

GGGAAGTCCT (SEQ ID NO:16); TGGCTGGGTA (SEQ ID NO:17);
CAATTAAGGA (SEQ ID NO:18); TATGTCCCTT (SEQ ID NO:19);
GCCTCTCATT (SEQ ID NO:20); TAGTTTTCCA (SEQ ID NO:21);
GTGCAATCCC (SEQ ID NO:22); and TCCCTCACAG (SEQ ID NO:23).

In some embodiments, a composition contains two or more differently labeled genotyping

oligonucleotides for simultaneously probing the identity of nucleotides at two or more polymorphic sites. It is also contemplated that primer compositions may contain two or more sets of allele-specific primer pairs to allow simultaneous targeting and amplification of two or more regions containing a polymorphic site.

SLC26A2 genotyping oligonucleotides of the invention may also be immobilized on or synthesized on a solid surface such as a microchip, bead, or glass slide (see, e.g., WO 98/20020 and WO 98/20019). Such immobilized genotyping oligonucleotides may be used in a variety of polymorphism detection assays, including but not limited to probe hybridization and polymerase extension assays. Immobilized SLC26A2 genotyping oligonucleotides of the invention may comprise an ordered array of oligonucleotides designed to rapidly screen a DNA sample for polymorphisms in multiple genes at the same time.

In another embodiment, the invention provides a kit comprising at least two genotyping oligonucleotides packaged in separate containers. The kit may also contain other components such as hybridization buffer (where the oligonucleotides are to be used as a probe) packaged in a separate container. Alternatively, where the oligonucleotides are to be used to amplify a target region, the kit may contain, packaged in separate containers, a polymerase and a reaction buffer optimized for primer extension mediated by the polymerase, such as PCR.

The above described oligonucleotide compositions and kits are useful in methods for genotyping and/or haplotyping the SLC26A2 gene in an individual. As used herein, the terms "SLC26A2 genotype" and "SLC26A2 haplotype" mean the genotype or haplotype contains the nucleotide pair or nucleotide, respectively, that is present at one or more of the novel polymorphic sites described herein and may optionally also include the nucleotide pair or nucleotide present at one or more additional polymorphic sites in the SLC26A2 gene. The additional polymorphic sites may be currently known polymorphic sites or sites that are subsequently discovered.

One embodiment of the genotyping method involves isolating from the individual a nucleic acid sample comprising the two copies of the SLC26A2 gene, or a fragment thereof, that are present in the individual, and determining the identity of the nucleotide pair at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3 and PS5 in the two copies to assign a SLC26A2 genotype to the individual. As will be readily understood by the skilled artisan, the two "copies" of a gene in an individual may be the same allele or may be different alleles. In a preferred embodiment of the genotyping method, the identity of the nucleotide pair at PS4 is also determined. In a particularly preferred embodiment, the genotyping method comprises determining the identity of the nucleotide pair at each of PS1-PS5.

Typically, the nucleic acid sample is isolated from a biological sample taken from the individual, such as a blood sample or tissue sample. Suitable tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The nucleic acid sample may be comprised of genomic DNA, mRNA, or cDNA and, in the latter two cases, the biological sample

must be obtained from a tissue in which the SLC26A2 gene is expressed. Furthermore it will be understood by the skilled artisan that mRNA or cDNA preparations would not be used to detect polymorphisms located in introns or in 5′ and 3′ untranslated regions. If a SLC26A2 gene fragment is isolated, it must contain the polymorphic site(s) to be genotyped.

One embodiment of the haplotyping method comprises isolating from the individual a nucleic acid sample containing only one of the two copies of the SLC26A2 gene, or a fragment thereof, that is present in the individual and determining in that copy the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3 and PS5 in that copy to assign a SLC26A2 haplotype to the individual. The nucleic acid may be isolated using any method capable of separating the two copies of the SLC26A2 gene or fragment such as one of the methods described above for preparing SLC26A2 isogenes, with targeted *in vivo* cloning being the preferred approach. As will be readily appreciated by those skilled in the art, any individual clone will only provide haplotype information on one of the two SLC26A2 gene copies present in an individual. If haplotype information is desired for the individual's other copy, additional SLC26A2 clones will need to be examined. Typically, at least five clones should be examined to have more than a 90% probability of haplotyping both copies of the SLC26A2 gene in an individual. In some embodiments, the haplotyping method also comprises identifying the nucleotide at PS4. In a particularly preferred embodiment, the nucleotide at each of PS1-PS5 is identified.

In another embodiment, the haplotyping method comprises determining whether an individual has one or more of the SLC26A2 haplotypes shown in Table 5. This can be accomplished by identifying, for one or both copies of the individual's SLC26A2 gene, the phased sequence of nucleotides present at each of PS1-PS5. The present invention also contemplates that typically only a subset of PS1-PS5 will need to be directly examined to assign to an individual one or more of the haplotypes shown in Table 5. This is because at least one polymorphic site in a gene is frequently in strong linkage disequilibrium with one or more other polymorphic sites in that gene (Drysdale, CM et al. 2000 *PNAS* 97:10483-10488; Rieder MJ et al. 1999 *Nature Genetics* 22:59-62). Two sites are said to be in linkage disequilibrium if the presence of a particular variant at one site enhances the predictability of another variant at the second site (Stephens, JC 1999, *Mol. Diag.* 4:309-317). Techniques for determining whether any two polymorphic sites are in linkage disequilibrium are well-known in the art (Weir B.S. 1996 *Genetic Data Analysis II*, Sinauer Associates, Inc. Publishers, Sunderland, MA).

In a preferred embodiment, a SLC26A2 haplotype pair is determined for an individual by identifying the phased sequence of nucleotides at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3 and PS5 in each copy of the SLC26A2 gene that is present in the individual. In a particularly preferred embodiment, the haplotyping method comprises identifying the phased sequence of nucleotides at each of PS1-PS5 in each copy of the SLC26A2 gene. When haplotyping both copies of the gene, the identifying step is preferably performed with each copy of the

gene being placed in separate containers. However, it is also envisioned that if the two copies are labeled with different tags, or are otherwise separately distinguishable or identifiable, it could be possible in some cases to perform the method in the same container. For example, if first and second copies of the gene are labeled with different first and second fluorescent dyes, respectively, and an allele-specific oligonucleotide labeled with yet a third different fluorescent dye is used to assay the polymorphic site(s), then detecting a combination of the first and third dyes would identify the polymorphism in the first gene copy while detecting a combination of the second and third dyes would identify the polymorphism in the second gene copy.

In both the genotyping and haplotyping methods, the identity of a nucleotide (or nucleotide pair) at a polymorphic site(s) may be determined by amplifying a target region(s) containing the polymorphic site(s) directly from one or both copies of the SLC26A2 gene, or a fragment thereof, and the sequence of the amplified region(s) determined by conventional methods. It will be readily appreciated by the skilled artisan that only one nucleotide will be detected at a polymorphic site in individuals who are homozygous at that site, while two different nucleotides will be detected if the individual is heterozygous for that site. The polymorphism may be identified directly, known as positive-type identification, or by inference, referred to as negative-type identification. For example, where a SNP is known to be guanine and cytosine in a reference population, a site may be positively determined to be either guanine or cytosine for an individual homozygous at that site, or both guanine and cytosine, if the individual is heterozygous at that site. Alternatively, the site may be negatively determined to be not guanine (and thus cytosine/cytosine) or not cytosine (and thus guanine/guanine).

The target region(s) may be amplified using any oligonucleotide-directed amplification method, including but not limited to polymerase chain reaction (PCR) (U.S. Patent No. 4,965,188), ligase chain reaction (LCR) (Barany et al., *Proc. Natl. Acad. Sci. USA* 88:189-193, 1991; WO90/01069), and oligonucleotide ligation assay (OLA) (Landegren et al., *Science* 241:1077-1080, 1988).

Other known nucleic acid amplification procedures may be used to amplify the target region including transcription-based amplification systems (U.S. Patent No. 5,130,238; EP 329,822; U.S. Patent No. 5,169,766, WO89/06700) and isothermal methods (Walker et al., *Proc. Natl. Acad. Sci. USA* 89:392-396, 1992).

A polymorphism in the target region may also be assayed before or after amplification using one of several hybridization-based methods known in the art. Typically, allele-specific oligonucleotides are utilized in performing such methods. The allele-specific oligonucleotides may be used as differently labeled probe pairs, with one member of the pair showing a perfect match to one variant of a target sequence and the other member showing a perfect match to a different variant. In some embodiments, more than one polymorphic site may be detected at once using a set of allele-specific oligonucleotides or oligonucleotide pairs. Preferably, the members of the set have melting temperatures within 5°C, and more preferably within 2°C, of each other when hybridizing to each of

the polymorphic sites being detected.

Hybridization of an allele-specific oligonucleotide to a target polynucleotide may be performed with both entities in solution, or such hybridization may be performed when either the oligonucleotide or the target polynucleotide is covalently or noncovalently affixed to a solid support. Attachment may be mediated, for example, by antibody-antigen interactions, poly-L-Lys, streptavidin or avidin-biotin, salt bridges, hydrophobic interactions, chemical linkages, UV cross-linking baking, etc. Allele-specific oligonucleotides may be synthesized directly on the solid support or attached to the solid support subsequent to synthesis. Solid-supports suitable for use in detection methods of the invention include substrates made of silicon, glass, plastic, paper and the like, which may be formed, for example, into wells (as in 96-well plates), slides, sheets, membranes, fibers, chips, dishes, and beads. The solid support may be treated, coated or derivatized to facilitate the immobilization of the allele-specific oligonucleotide or target nucleic acid.

The genotype or haplotype for the SLC26A2 gene of an individual may also be determined by hybridization of a nucleic acid sample containing one or both copies of the gene, or fragment(s) thereof, to nucleic acid arrays and subarrays such as described in WO 95/11995. The arrays would contain a battery of allele-specific oligonucleotides representing each of the polymorphic sites to be included in the genotype or haplotype.

The identity of polymorphisms may also be determined using a mismatch detection technique, including but not limited to the RNase protection method using riboprobes (Winter et al., *Proc. Natl. Acad. Sci. USA* 82:7575, 1985; Meyers et al., *Science* 230:1242, 1985) and proteins which recognize nucleotide mismatches, such as the E. coli mutS protein (Modrich, P. *Ann. Rev. Genet.* 25:229-253, 1991). Alternatively, variant alleles can be identified by single strand conformation polymorphism (SSCP) analysis (Orita et al., *Genomics* 5:874-879, 1989; Humphries et al., in Molecular Diagnosis of Genetic Diseases, R. Elles, ed., pp. 321-340, 1996) or denaturing gradient gel electrophoresis (DGGE) (Wartell et al., *Nucl. Acids Res.* 18:2699-2706, 1990; Sheffield et al., *Proc. Natl. Acad. Sci. USA* 86:232-236, 1989).

A polymerase-mediated primer extension method may also be used to identify the polymorphism(s). Several such methods have been described in the patent and scientific literature and include the "Genetic Bit Analysis" method (WO92/15712) and the ligase/polymerase mediated genetic bit analysis (U.S. Patent 5,679,524. Related methods are disclosed in WO91/02087, WO90/09455, WO95/17676, U.S. Patent Nos. 5,302,509, and 5,945,283. Extended primers containing a polymorphism may be detected by mass spectrometry as described in U.S. Patent No. 5,605,798. Another primer extension method is allele-specific PCR (Ruaño et al., *Nucl. Acids Res.* 17:8392, 1989; Ruaño et al., *Nucl. Acids Res.* 19, 6877-6882, 1991; WO 93/22456; Turki et al., *J. Clin. Invest.* 95:1635-1641, 1995). In addition, multiple polymorphic sites may be investigated by simultaneously amplifying multiple regions of the nucleic acid using sets of allele-specific primers as described in Wallace et al. (WO89/10414).

In addition, the identity of the allele(s) present at any of the novel polymorphic sites described herein may be indirectly determined by genotyping another polymorphic site that is in linkage disequilibrium with the polymorphic site that is of interest. Polymorphic sites in linkage disequilibrium with the presently disclosed polymorphic sites may be located in regions of the gene or in other genomic regions not examined herein. Genotyping of a polymorphic site in linkage disequilibrium with the novel polymorphic sites described herein may be performed by, but is not limited to, any of the above-mentioned methods for detecting the identity of the allele at a polymorphic site.

In another aspect of the invention, an individual's SLC26A2 haplotype pair is predicted from its SLC26A2 genotype using information on haplotype pairs known to exist in a reference population. In its broadest embodiment, the haplotyping prediction method comprises identifying a SLC26A2 genotype for the individual at two or more SLC26A2 polymorphic sites described herein, enumerating all possible haplotype pairs which are consistent with the genotype, accessing data containing SLC26A2 haplotype pairs identified in a reference population, and assigning a haplotype pair to the individual that is consistent with the data. In one embodiment, the reference haplotype pairs include the SLC26A2 haplotype pairs shown in Table 4.

Generally, the reference population should be composed of randomly-selected individuals representing the major ethnogeographic groups of the world. A preferred reference population for use in the methods of the present invention comprises an approximately equal number of individuals from Caucasian, African-descent, Asian and Hispanic-Latino population groups with the minimum number of each group being chosen based on how rare a haplotype one wants to be guaranteed to see. For example, if one wants to have a q% chance of not missing a haplotype that exists in the population at a p% frequency of occurring in the reference population, the number of individuals (n) who must be sampled is given by $2n=\log(1-q)/\log(1-p)$ where p and q are expressed as fractions. A preferred reference population allows the detection of any haplotype whose frequency is at least 10% with about 99% certainty and comprises about 20 unrelated individuals from each of the four population groups named above. A particularly preferred reference population includes a 3-generation family representing one or more of the four population groups to serve as controls for checking quality of haplotyping procedures.

In a preferred embodiment, the haplotype frequency data for each ethnogeographic group is examined to determine whether it is consistent with Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium (D.L. Hartl et al., Principles of Population Genomics, Sinauer Associates (Sunderland, MA), 3rd Ed., 1997) postulates that the frequency of finding the haplotype pair H_1 / H_2 is equal to $p_{H-W}(H_1 / H_2) = 2p(H_1)p(H_2)$ if $H_1 \neq H_2$ and $p_{H-W}(H_1 / H_2) = p(H_1)p(H_2)$ if $H_1 = H_2$. A statistically significant difference between the observed and expected haplotype frequencies could be due to one or more factors including significant inbreeding in the population group, strong selective

pressure on the gene, sampling bias, and/or errors in the genotyping process. If large deviations from Hardy-Weinberg equilibrium are observed in an ethnogeographic group, the number of individuals in that group can be increased to see if the deviation is due to a sampling bias. If a larger sample size does not reduce the difference between observed and expected haplotype pair frequencies, then one may wish to consider haplotyping the individual using a direct haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), single molecule dilution, or allele-specific long-range PCR (Michalotos-Beloin et al., *Nucleic Acids Res.* 24:4841-4843, 1996).

In one embodiment of this method for predicting a SLC26A2 haplotype pair for an individual, the assigning step involves performing the following analysis. First, each of the possible haplotype pairs is compared to the haplotype pairs in the reference population. Generally, only one of the haplotype pairs in the reference population matches a possible haplotype pair and that pair is assigned to the individual. Occasionally, only one haplotype represented in the reference haplotype pairs is consistent with a possible haplotype pair for an individual, and in such cases the individual is assigned a haplotype pair containing this known haplotype and a new haplotype derived by subtracting the known haplotype from the possible haplotype pair. Alternatively, the haplotype pair in an individual may be predicted from the individual's genotype for that gene using reported methods (e.g., Clark et al. 1990 *Mol Bio Evol* 7:111-22) or through a commercial haplotyping service such as offered by Genaissance Pharmaceuticals, Inc. (New Haven, CT). In rare cases, either no haplotypes in the reference population are consistent with the possible haplotype pairs, or alternatively, multiple reference haplotype pairs are consistent with the possible haplotype pairs. In such cases, the individual is preferably haplotyped using a direct molecular haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), SMD, or allele-specific long-range PCR (Michalotos-Beloin et al., *supra*). A preferred process for predicting SLC26A2 haplotype pairs from SLC26A2 genotypes is described in U.S. Provisional Application Serial No. 60/198,340 and the corresponding International Application, PCT/US01/12831.

The invention also provides a method for determining the frequency of a SLC26A2 genotype, haplotype, or haplotype pair in a population. The method comprises, for each member of the population, determining the genotype or the haplotype pair for the novel SLC26A2 polymorphic sites described herein, and calculating the frequency any particular genotype, haplotype, or haplotype pair is found in the population. The population may be a reference population, a family population, a same sex population, a population group, or a trait population (e.g., a group of individuals exhibiting a trait of interest such as a medical condition or response to a therapeutic treatment).

In another aspect of the invention, frequency data for SLC26A2 genotypes, haplotypes, and/or haplotype pairs are determined in a reference population and used in a method for identifying an association between a trait and a SLC26A2 genotype, haplotype, or haplotype pair. The trait may be any detectable phenotype, including but not limited to susceptibility to a disease or response to a treatment. The method involves obtaining data on the frequency of the genotype(s), haplotype(s), or

haplotype pair(s) of interest in a reference population as well as in a population exhibiting the trait. Frequency data for one or both of the reference and trait populations may be obtained by genotyping or haplotyping each individual in the populations using one of the methods described above. The haplotypes for the trait population may be determined directly or, alternatively, by the predictive genotype to haplotype approach described above. In another embodiment, the frequency data for the reference and/or trait populations is obtained by accessing previously determined frequency data, which may be in written or electronic form. For example, the frequency data may be present in a database that is accessible by a computer. Once the frequency data is obtained, the frequencies of the genotype(s), haplotype(s), or haplotype pair(s) of interest in the reference and trait populations are compared. In a preferred embodiment, the frequencies of all genotypes, haplotypes, and/or haplotype pairs observed in the populations are compared. If a particular SLC26A2 genotype, haplotype, or haplotype pair is more frequent in the trait population than in the reference population at a statistically significant amount, then the trait is predicted to be associated with that SLC26A2 genotype, haplotype or haplotype pair. Preferably, the SLC26A2 genotype, haplotype, or haplotype pair being compared in the trait and reference populations is selected from the full-genotypes and full-haplotypes shown in Tables 4 and 5, or from sub-genotypes and sub-haplotypes derived from these genotypes and haplotypes.

In a preferred embodiment of the method, the trait of interest is a clinical response exhibited by a patient to some therapeutic treatment, for example, response to a drug targeting SLC26A2 or response to a therapeutic treatment for a medical condition. As used herein, "medical condition" includes but is not limited to any condition or disease manifested as one or more physical and/or psychological symptoms for which treatment is desirable, and includes previously and newly identified diseases and other disorders. As used herein the term "clinical response" means any or all of the following: a quantitative measure of the response, no response, and adverse response (i.e., side effects).

In order to deduce a correlation between clinical response to a treatment and a SLC26A2 genotype, haplotype, or haplotype pair, it is necessary to obtain data on the clinical responses exhibited by a population of individuals who received the treatment, hereinafter the "clinical population". This clinical data may be obtained by analyzing the results of a clinical trial that has already been run and/or the clinical data may be obtained by designing and carrying out one or more new clinical trials. As used herein, the term "clinical trial" means any research study designed to collect clinical data on responses to a particular treatment, and includes but is not limited to phase I, phase II and phase III clinical trials. Standard methods are used to define the patient population and to enroll subjects.

It is preferred that the individuals included in the clinical population have been graded for the existence of the medical condition of interest. This is important in cases where the symptom(s) being presented by the patients can be caused by more than one underlying condition, and where treatment

of the underlying conditions are not the same. An example of this would be where patients experience breathing difficulties that are due to either asthma or respiratory infections. If both sets were treated with an asthma medication, there would be a spurious group of apparent non-responders that did not actually have asthma. These people would affect the ability to detect any correlation between haplotype and treatment outcome. This grading of potential patients could employ a standard physical exam or one or more lab tests. Alternatively, grading of patients could use haplotyping for situations where there is a strong correlation between haplotype pair and disease susceptibility or severity.

The therapeutic treatment of interest is administered to each individual in the trial population and each individual's response to the treatment is measured using one or more predetermined criteria. It is contemplated that in many cases, the trial population will exhibit a range of responses and that the investigator will choose the number of responder groups (e.g., low, medium, high) made up by the various responses. In addition, the SLC26A2 gene for each individual in the trial population is genotyped and/or haplotyped, which may be done before or after administering the treatment.

After both the clinical and polymorphism data have been obtained, correlations between individual response and SLC26A2 genotype or haplotype content are created. Correlations may be produced in several ways. In one method, individuals are grouped by their SLC26A2 genotype or haplotype (or haplotype pair) (also referred to as a polymorphism group), and then the averages and standard deviations of clinical responses exhibited by the members of each polymorphism group are calculated.

These results are then analyzed to determine if any observed variation in clinical response between polymorphism groups is statistically significant. Statistical analysis methods which may be used are described in L.D. Fisher and G. vanBelle, "Biostatistics: A Methodology for the Health Sciences", Wiley-Interscience (New York) 1993. This analysis may also include a regression calculation of which polymorphic sites in the SLC26A2 gene give the most significant contribution to the differences in phenotype. One regression model useful in the invention is described in PCT Application Serial No. PCT/US00/17540, entitled "Methods for Obtaining and Using Haplotype Data".

A second method for finding correlations between SLC26A2 haplotype content and clinical responses uses predictive models based on error-minimizing optimization algorithms. One of many possible optimization algorithms is a genetic algorithm (R. Judson, "Genetic Algorithms and Their Uses in Chemistry" in Reviews in Computational Chemistry, Vol. 10, pp. 1-73, K. B. Lipkowitz and D. B. Boyd, eds. (VCH Publishers, New York, 1997). Simulated annealing (Press et al., "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press (Cambridge) 1992, Ch. 10), neural networks (E. Rich and K. Knight, "Artificial Intelligence", 2nd Edition (McGraw-Hill, New York, 1991, Ch. 18), standard gradient descent methods (Press et al., *supra*, Ch. 10), or other global or local optimization approaches (see discussion in Judson, *supra*) could also be used. Preferably, the correlation is found using a genetic algorithm approach as described in PCT Application Serial No.

PCT/US00/17540.

Correlations may also be analyzed using analysis of variation (ANOVA) techniques to determine how much of the variation in the clinical data is explained by different subsets of the polymorphic sites in the SLC26A2 gene. As described in PCT Application Serial No. PCT/US00/17540, ANOVA is used to test hypotheses about whether a response variable is caused by or correlated with one or more traits or variables that can be measured (Fisher and vanBelle, *supra*, Ch. 10).

From the analyses described above, a mathematical model may be readily constructed by the skilled artisan that predicts clinical response as a function of SLC26A2 genotype or haplotype content. Preferably, the model is validated in one or more follow-up clinical trials designed to test the model.

The identification of an association between a clinical response and a genotype or haplotype (or haplotype pair) for the SLC26A2 gene may be the basis for designing a diagnostic method to determine those individuals who will or will not respond to the treatment, or alternatively, will respond at a lower level and thus may require more treatment, i.e., a greater dose of a drug. The diagnostic method may take one of several forms: for example, a direct DNA test (i.e., genotyping or haplotyping one or more of the polymorphic sites in the SLC26A2 gene), a serological test, or a physical exam measurement. The only requirement is that there be a good correlation between the diagnostic test results and the underlying SLC26A2 genotype or haplotype that is in turn correlated with the clinical response. In a preferred embodiment, this diagnostic method uses the predictive haplotyping method described above.

In another embodiment, the invention provides an isolated polynucleotide comprising a polymorphic variant of the SLC26A2 gene or a fragment of the gene which contains at least one of the novel polymorphic sites described herein. The nucleotide sequence of a variant SLC26A2 gene is identical to the reference genomic sequence for those portions of the gene examined, as described in the Examples below, except that it comprises a different nucleotide at one or more of the novel polymorphic sites PS1, PS2, PS3 and PS5, and may also comprise an additional polymorphism of thymine at PS4. Similarly, the nucleotide sequence of a variant fragment of the SLC26A2 gene is identical to the corresponding portion of the reference sequence except for having a different nucleotide at one or more of the novel polymorphic sites described herein. Thus, the invention specifically does not include polynucleotides comprising a nucleotide sequence identical to the reference sequence of the SLC26A2 gene, which is defined by haplotype 3, (or other reported SLC26A2 sequences) or to portions of the reference sequence (or other reported SLC26A2 sequences), except for genotyping oligonucleotides as described above.

The location of a polymorphism in a variant gene or fragment is identified by aligning its sequence against SEQ ID NO:1. The polymorphism is selected from the group consisting of adenine at PS1, guanine at PS2, adenine at PS3 and thymine at PS5. In a preferred embodiment, the polymorphic variant comprises a naturally-occurring isogene of the SLC26A2 gene which is defined

by any one of haplotypes 1- 2 and 4 - 5 shown in Table 5 below.

Polymorphic variants of the invention may be prepared by isolating a clone containing the SLC26A2 gene from a human genomic library. The clone may be sequenced to determine the identity of the nucleotides at the novel polymorphic sites described herein. Any particular variant claimed herein could be prepared from this clone by performing *in vitro* mutagenesis using procedures well-known in the art.

SLC26A2 isogenes may be isolated using any method that allows separation of the two "copies" of the SLC26A2 gene present in an individual, which, as readily understood by the skilled artisan, may be the same allele or different alleles. Separation methods include targeted *in vivo* cloning (TIVC) in yeast as described in WO 98/01573, U.S. Patent No. 5,866,404, and U.S. Patent No. 5,972,614. Another method, which is described in U.S. Patent No. 5,972,614, uses an allele specific oligonucleotide in combination with primer extension and exonuclease degradation to generate hemizygous DNA targets. Yet other methods are single molecule dilution (SMD) as described in Ruaño et al., *Proc. Natl. Acad. Sci.* 87:6296-6300, 1990; and allele specific PCR (Ruaño et al., 1989, *supra*; Ruaño et al., 1991, *supra*; Michalatos-Beloin et al., *supra*).

The invention also provides SLC26A2 genome anthologies, which are collections of SLC26A2 isogenes found in a given population. The population may be any group of at least two individuals, including but not limited to a reference population, a population group, a family population, a clinical population, and a same sex population. A SLC26A2 genome anthology may comprise individual SLC26A2 isogenes stored in separate containers such as microtest tubes, separate wells of a microtitre plate and the like. Alternatively, two or more groups of the SLC26A2 isogenes in the anthology may be stored in separate containers. Individual isogenes or groups of isogenes in a genome anthology may be stored in any convenient and stable form, including but not limited to in buffered solutions, as DNA precipitates, freeze-dried preparations and the like. A preferred SLC26A2 genome anthology of the invention comprises a set of isogenes defined by the haplotypes shown in Table 5 below.

An isolated polynucleotide containing a polymorphic variant nucleotide sequence of the invention may be operably linked to one or more expression regulatory elements in a recombinant expression vector capable of being propagated and expressing the encoded SLC26A2 protein in a prokaryotic or a eukaryotic host cell. Examples of expression regulatory elements which may be used include, but are not limited to, the lac system, operator and promoter regions of phage lambda, yeast promoters, and promoters derived from vaccinia virus, adenovirus, retroviruses, or SV40. Other regulatory elements include, but are not limited to, appropriate leader sequences, termination codons, polyadenylation signals, and other sequences required for the appropriate transcription and subsequent translation of the nucleic acid sequence in a given host cell. Of course, the correct combinations of expression regulatory elements will depend on the host system used. In addition, it is understood that the expression vector contains any additional elements necessary for its transfer to and subsequent

replication in the host cell. Examples of such elements include, but are not limited to, origins of replication and selectable markers. Such expression vectors are commercially available or are readily constructed using methods known to those in the art (e.g., F. Ausubel et al., 1987, in "Current Protocols in Molecular Biology", John Wiley and Sons, New York, New York). Host cells which may be used to express the variant SLC26A2 sequences of the invention include, but are not limited to, eukaryotic and mammalian cells, such as animal, plant, insect and yeast cells, and prokaryotic cells, such as E. coli, or algal cells as known in the art. The recombinant expression vector may be introduced into the host cell using any method known to those in the art including, but not limited to, microinjection, electroporation, particle bombardment, transduction, and transfection using DEAE-dextran, lipofection, or calcium phosphate (see e.g., Sambrook et al. (1989) in "Molecular Cloning. A Laboratory Manual", Cold Spring Harbor Press, Plainview, New York). In a preferred aspect, eukaryotic expression vectors that function in eukaryotic cells, and preferably mammalian cells, are used. Non-limiting examples of such vectors include vaccinia virus vectors, adenovirus vectors, herpes virus vectors, and baculovirus transfer vectors. Preferred eukaryotic cell lines include COS cells, CHO cells, HeLa cells, NIH/3T3 cells, and embryonic stem cells (Thomson, J. A. et al., 1998 *Science* 282:1145-1147). Particularly preferred host cells are mammalian cells.

As will be readily recognized by the skilled artisan, expression of polymorphic variants of the SLC26A2 gene will produce SLC26A2 mRNAs varying from each other at any polymorphic site retained in the spliced and processed mRNA molecules. These mRNAs can be used for the preparation of a SLC26A2 cDNA comprising a nucleotide sequence which is a polymorphic variant of the SLC26A2 reference coding sequence shown in Figure 2. Thus, the invention also provides SLC26A2 mRNAs and corresponding cDNAs which comprise a nucleotide sequence that is identical to SEQ ID NO:2 (Fig. 2), or its corresponding RNA sequence, except for having one or more polymorphisms selected from the group consisting of adenine at a position corresponding to nucleotide 1046 and thymine at a position corresponding to nucleotide 2065, and may also comprise an additional polymorphism of thymine at a position corresponding to nucleotide 1721. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a SLC26A2 isogene defined by haplotypes 1, 2, 4 and 5. Fragments of these variant mRNAs and cDNAs are included in the scope of the invention, provided they contain the novel polymorphisms described herein. The invention specifically excludes polynucleotides identical to previously identified and characterized SLC26A2 cDNAs and fragments thereof. Polynucleotides comprising a variant RNA or DNA sequence may be isolated from a biological sample using well-known molecular biological procedures or may be chemically synthesized.

As used herein, a polymorphic variant of a SLC26A2 gene fragment comprises at least one novel polymorphism identified herein and has a length of at least 10 nucleotides and may range up to the full length of the gene. Preferably, such fragments are between 100 and 3000 nucleotides in length, and more preferably between 200 and 2000 nucleotides in length, and most preferably between

500 and 1000 nucleotides in length.

In describing the SLC26A2 polymorphic sites identified herein, reference is made to the sense strand of the gene for convenience. However, as recognized by the skilled artisan, nucleic acid molecules containing the SLC26A2 gene may be complementary double stranded molecules and thus reference to a particular site on the sense strand refers as well to the corresponding site on the complementary antisense strand. Thus, reference may be made to the same polymorphic site on either strand and an oligonucleotide may be designed to hybridize specifically to either strand at a target region containing the polymorphic site. Thus, the invention also includes single-stranded polynucleotides which are complementary to the sense strand of the SLC26A2 genomic variants described herein.

Polynucleotides comprising a polymorphic gene variant or fragment may be useful for therapeutic purposes. For example, where a patient could benefit from expression, or increased expression, of a particular SLC26A2 protein isoform, an expression vector encoding the isoform may be administered to the patient. The patient may be one who lacks the SLC26A2 isogene encoding that isoform or may already have at least one copy of that isogene.

In other situations, it may be desirable to decrease or block expression of a particular SLC26A2 isogene. Expression of a SLC26A2 isogene may be turned off by transforming a targeted organ, tissue or cell population with an expression vector that expresses high levels of untranslatable mRNA for the isogene. Alternatively, oligonucleotides directed against the regulatory regions (e.g., promoter, introns, enhancers, 3′ untranslated region) of the isogene may block transcription. Oligonucleotides targeting the transcription initiation site, e.g., between positions –10 and +10 from the start site are preferred. Similarly, inhibition of transcription can be achieved using oligonucleotides that base-pair with region(s) of the isogene DNA to form triplex DNA (see e.g., Gee et al. in Huber, B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, N.Y., 1994). Antisense oligonucleotides may also be designed to block translation of SLC26A2 mRNA transcribed from a particular isogene. It is also contemplated that ribozymes may be designed that can catalyze the specific cleavage of SLC26A2 mRNA transcribed from a particular isogene.

The oligonucleotides may be delivered to a target cell or tissue by expression from a vector introduced into the cell or tissue *in vivo* or *ex vivo*. Alternatively, the oligonucleotides may be formulated as a pharmaceutical composition for administration to the patient. Oligoribonucleotides and/or oligodeoxynucleotides intended for use as antisense oligonucleotides may be modified to increase stability and half-life. Possible modifications include, but are not limited to phosphorothioate or 2′ O-methyl linkages, and the inclusion of nontraditional bases such as inosine and queosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytosine, guanine, thymine, and uracil which are not as easily recognized by endogenous nucleases.

The invention also provides an isolated polypeptide comprising a polymorphic variant of the

reference SLC26A2 amino acid sequence shown in Figure 3. The location of a variant amino acid in a SLC26A2 polypeptide or fragment of the invention is identified by aligning its sequence against SEQ ID NO:3 (Fig. 3). A SLC26A2 protein variant of the invention comprises an amino acid sequence identical to SEQ ID NO:3 except for having one or more variant amino acids selected from the group consisting oftyrosine at a position corresponding to amino acid position 349 and serine at a position corresponding to amino acid position 689, and may also comprise an additional variant amino acid of isoleucine at a position corresponding to amino acid position 574. The invention specifically excludes amino acid sequences identical to those previously identified for SLC26A2, including SEQ ID NO:3, and previously described fragments thereof. SLC26A2 protein variants included within the invention comprise all amino acid sequences based on SEQ ID NO:3 and having the combination of amino acid variations described in Table 2 below. In preferred embodiments, a SLC26A2 protein variant of the invention is encoded by an isogene defined by one of the observed haplotypes shown in Table 5.

Table 2. Novel Polymorphic Variants of SLC26A2

Polymorphic Variant Number	Amino Acid Position and Identities		
	349	574	689
1	F	T	S
2	F	I	S
3	Y	T	T
4	Y	T	S
5	Y	I	T
6	Y	I	S

The invention also includes SLC26A2 peptide variants, which are any fragments of a SLC26A2 protein variant that contain one or more of the amino acid variations shown in Table 2 A SLC26A2 peptide variant is at least 6 amino acids in length and is preferably any number between 6 and 30 amino acids long, more preferably between 10 and 25, and most preferably between 15 and 20 amino acids long. Such SLC26A2 peptide variants may be useful as antigens to generate antibodies specific for one of the above SLC26A2 isoforms. In addition, the SLC26A2 peptide variants may be useful in drug screening assays.

A SLC26A2 variant protein or peptide of the invention may be prepared by chemical synthesis or by expressing one of the variant SLC26A2 genomic and cDNA sequences as described above. Alternatively, the SLC26A2 protein variant may be isolated from a biological sample of an individual having a SLC26A2 isogene which encodes the variant protein. Where the sample contains two different SLC26A2 isoforms (i.e., the individual has different SLC26A2 isogenes), a particular SLC26A2 isoform of the invention can be isolated by immunoaffinity chromatography using an antibody which specifically binds to that particular SLC26A2 isoform but does not bind to the other SLC26A2 isoform.

The expressed or isolated SLC26A2 protein may be detected by methods known in the art, including Coomassie blue staining, silver staining, and Western blot analysis using antibodies specific

for the isoform of the SLC26A2 protein as discussed further below. SLC26A2 variant proteins can be purified by standard protein purification procedures known in the art, including differential precipitation, molecular sieve chromatography, ion-exchange chromatography, isoelectric focusing, gel electrophoresis, affinity and immunoaffinity chromatography and the like. (Ausubel et. al., 1987, In Current Protocols in Molecular Biology John Wiley and Sons, New York, New York). In the case of immunoaffinity chromatography, antibodies specific for a particular polymorphic variant may be used.

A polymorphic variant SLC26A2 gene of the invention may also be fused in frame with a heterologous sequence to encode a chimeric SLC26A2 protein. The non-SLC26A2 portion of the chimeric protein may be recognized by a commercially available antibody. In addition, the chimeric protein may also be engineered to contain a cleavage site located between the SLC26A2 and non-SLC26A2 portions so that the SLC26A2 protein may be cleaved and purified away from the non-SLC26A2 portion.

An additional embodiment of the invention relates to using a novel SLC26A2 protein isoform in any of a variety of drug screening assays. Such screening assays may be performed to identify agents that bind specifically to all known SLC26A2 protein isoforms or to only a subset of one or more of these isoforms. The agents may be from chemical compound libraries, peptide libraries and the like. The SLC26A2 protein or peptide variant may be free in solution or affixed to a solid support. In one embodiment, high throughput screening of compounds for binding to a SLC26A2 variant may be accomplished using the method described in PCT application WO84/03565, in which large numbers of test compounds are synthesized on a solid substrate, such as plastic pins or some other surface, contacted with the SLC26A2 protein(s) of interest and then washed. Bound SLC26A2 protein(s) are then detected using methods well-known in the art.

In another embodiment, a novel SLC26A2 protein isoform may be used in assays to measure the binding affinities of one or more candidate drugs targeting the SLC26A2 protein.

In yet another embodiment, when a particular SLC26A2 haplotype or group of SLC26A2 haplotypes encodes a SLC26A2 protein variant with an amino acid sequence distinct from that of SLC26A2 protein isoforms encoded by other SLC26A2 haplotypes, then detection of that particular SLC26A2 haplotype or group of SLC26A2 haplotypes may be accomplished by detecting expression of the encoded SLC26A2 protein variant using any of the methods described herein or otherwise commonly known to the skilled artisan.

In another embodiment, the invention provides antibodies specific for and immunoreactive with one or more of the novel SLC26A2 variant proteins described herein. The antibodies may be either monoclonal or polyclonal in origin. The SLC26A2 protein or peptide variant used to generate the antibodies may be from natural or recombinant sources or produced by chemical synthesis using synthesis techniques known in the art. If the SLC26A2 protein variant is of insufficient size to be antigenic, it may be conjugated, complexed, or otherwise covalently linked to a carrier molecule to

enhance the antigenicity of the peptide. Examples of carrier molecules, include, but are not limited to, albumins (e.g., human, bovine, fish, ovine), and keyhole limpet hemocyanin (Basic and Clinical Immunology, 1991, Eds. D.P. Stites, and A.I. Terr, Appleton and Lange, Norwalk Connecticut, San Mateo, California).

In one embodiment, an antibody specifically immunoreactive with one of the novel protein isoforms described herein is administered to an individual to neutralize activity of the SLC26A2 isoform expressed by that individual. The antibody may be formulated as a pharmaceutical composition which includes a pharmaceutically acceptable carrier.

Antibodies specific for and immunoreactive with one of the novel protein isoforms described herein may be used to immunoprecipitate the SLC26A2 protein variant from solution as well as react with SLC26A2 protein isoforms on Western or immunoblots of polyacrylamide gels on membrane supports or substrates. In another preferred embodiment, the antibodies will detect SLC26A2 protein isoforms in paraffin or frozen tissue sections, or in cells which have been fixed or unfixed and prepared on slides, coverslips, or the like, for use in immunocytochemical, immunohistochemical, and immunofluorescence techniques.

In another embodiment, an antibody specifically immunoreactive with one of the novel SLC26A2 protein variants described herein is used in immunoassays to detect this variant in biological samples. In this method, an antibody of the present invention is contacted with a biological sample and the formation of a complex between the SLC26A2 protein variant and the antibody is detected. As described, suitable immunoassays include radioimmunoassay, Western blot assay, immunofluorescent assay, enzyme linked immunoassay (ELISA), chemiluminescent assay, immunohistochemical assay, immunocytochemical assay, and the like (see, e.g., Principles and Practice of Immunoassay, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Press, New York, New York; Current Protocols in Molecular Biology, 1987, Eds. Ausubel et al., John Wiley and Sons, New York, New York). Standard techniques known in the art for ELISA are described in Methods in Immunodiagnosis, 2nd Ed., Eds. Rose and Bigazzi, John Wiley and Sons, New York 1980; and Campbell et al., 1984, Methods in Immunology, W.A. Benjamin, Inc.). Such assays may be direct, indirect, competitive, or noncompetitive as described in the art (see, e.g., Principles and Practice of Immunoassay, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Pres, NY, NY; and Oellirich, M., 1984, J. Clin. Chem. Clin. Biochem., 22:895-904). Proteins may be isolated from test specimens and biological samples by conventional methods, as described in Current Protocols in Molecular Biology, supra.

Exemplary antibody molecules for use in the detection and therapy methods of the present invention are intact immunoglobulin molecules, substantially intact immunoglobulin molecules, or those portions of immunoglobulin molecules that contain the antigen binding site. Polyclonal or monoclonal antibodies may be produced by methods conventionally known in the art (e.g., Kohler and Milstein, 1975, Nature, 256:495-497; Campbell Monoclonal Antibody Technology, the Production

and Characterization of Rodent and Human Hybridomas, 1985, In: Laboratory Techniques in Biochemistry and Molecular Biology, Eds. Burdon et al., Volume 13, Elsevier Science Publishers, Amsterdam). The antibodies or antigen binding fragments thereof may also be produced by genetic engineering. The technology for expression of both heavy and light chain genes in E. coli is the subject of PCT patent applications, publication number WO 901443, WO 901443 and WO 9014424 and in Huse et al., 1989, Science, 246:1275-1281. The antibodies may also be humanized (e.g., Queen, C. et al. 1989 Proc. Natl. Acad. Sci.USA 86;10029).

Effect(s) of the polymorphisms identified herein on expression of SLC26A2 may be investigated by preparing recombinant cells and/or nonhuman recombinant organisms, preferably recombinant animals, containing a polymorphic variant of the SLC26A2 gene. As used herein, "expression" includes but is not limited to one or more of the following: transcription of the gene into precursor mRNA; splicing and other processing of the precursor mRNA to produce mature mRNA; mRNA stability; translation of the mature mRNA into SLC26A2 protein (including codon usage and tRNA availability); and glycosylation and/or other modifications of the translation product, if required for proper expression and function.

To prepare a recombinant cell of the invention, the desired SLC26A2 isogene may be introduced into the cell in a vector such that the isogene remains extrachromosomal. In such a situation, the gene will be expressed by the cell from the extrachromosomal location. In a preferred embodiment, the SLC26A2 isogene is introduced into a cell in such a way that it recombines with the endogenous SLC26A2 gene present in the cell. Such recombination requires the occurrence of a double recombination event, thereby resulting in the desired SLC26A2 gene polymorphism. Vectors for the introduction of genes both for recombination and for extrachromosomal maintenance are known in the art, and any suitable vector or vector construct may be used in the invention. Methods such as electroporation, particle bombardment, calcium phosphate co-precipitation and viral transduction for introducing DNA into cells are known in the art; therefore, the choice of method may lie with the competence and preference of the skilled practitioner. Examples of cells into which the SLC26A2 isogene may be introduced include, but are not limited to, continuous culture cells, such as COS, NIH/3T3, and primary or culture cells of the relevant tissue type, i.e., they express the SLC26A2 isogene. Such recombinant cells can be used to compare the biological activities of the different protein variants.

Recombinant nonhuman organisms, i.e., transgenic animals, expressing a variant SLC26A2 gene are prepared using standard procedures known in the art. Preferably, a construct comprising the variant gene is introduced into a nonhuman animal or an ancestor of the animal at an embryonic stage, i.e., the one-cell stage, or generally not later than about the eight-cell stage. Transgenic animals carrying the constructs of the invention can be made by several methods known to those having skill in the art. One method involves transfecting into the embryo a retrovirus constructed to contain one or more insulator elements, a gene or genes of interest, and other components known to those skilled in

the art to provide a complete shuttle vector harboring the insulated gene(s) as a transgene, see e.g., U.S. Patent No. 5,610,053. Another method involves directly injecting a transgene into the embryo. A third method involves the use of embryonic stem cells. Examples of animals into which the SLC26A2 isogenes may be introduced include, but are not limited to, mice, rats, other rodents, and nonhuman primates (see "The Introduction of Foreign Genes into Mice" and the cited references therein, In: Recombinant DNA, Eds. J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller; W.H. Freeman and Company, New York, pages 254-272). Transgenic animals stably expressing a human SLC26A2 isogene and producing human SLC26A2 protein can be used as biological models for studying diseases related to abnormal SLC26A2 expression and/or activity, and for screening and assaying various candidate drugs, compounds, and treatment regimens to reduce the symptoms or effects of these diseases.

An additional embodiment of the invention relates to pharmaceutical compositions for treating disorders affected by expression or function of a novel SLC26A2 isogene described herein. The pharmaceutical composition may comprise any of the following active ingredients: a polynucleotide comprising one of these novel SLC26A2 isogenes; an antisense oligonucleotide directed against one of the novel SLC26A2 isogenes, a polynucleotide encoding such an antisense oligonucleotide, or another compound which inhibits expression of a novel SLC26A2 isogene described herein. Preferably, the composition contains the active ingredient in a therapeutically effective amount. By therapeutically effective amount is meant that one or more of the symptoms relating to disorders affected by expression or function of a novel SLC26A2 isogene is reduced and/or eliminated. The composition also comprises a pharmaceutically acceptable carrier, examples of which include, but are not limited to, saline, buffered saline, dextrose, and water. Those skilled in the art may employ a formulation most suitable for the active ingredient, whether it is a polynucleotide, oligonucleotide, protein, peptide or small molecule antagonist. The pharmaceutical composition may be administered alone or in combination with at least one other agent, such as a stabilizing compound. Administration of the pharmaceutical composition may be by any number of routes including, but not limited to oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, intradermal, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

For any composition, determination of the therapeutically effective dose of active ingredient and/or the appropriate route of administration is well within the capability of those skilled in the art. For example, the dose can be estimated initially either in cell culture assays or in animal models. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans. The exact dosage will be determined by the practitioner, in light of factors relating to the patient requiring treatment, including but not limited to severity of the disease state,

general health, age, weight and gender of the patient, diet, time and frequency of administration, other drugs being taken by the patient, and tolerance/response to the treatment.

Any or all analytical and mathematical operations involved in practicing the methods of the present invention may be implemented by a computer. In addition, the computer may execute a program that generates views (or screens) displayed on a display device and with which the user can interact to view and analyze large amounts of information relating to the SLC26A2 gene and its genomic variation, including chromosome location, gene structure, and gene family, gene expression data, polymorphism data, genetic sequence data, and clinical data population data (e.g., data on ethnogeographic origin, clinical responses, genotypes, and haplotypes for one or more populations). The SLC26A2 polymorphism data described herein may be stored as part of a relational database (e.g., an instance of an Oracle database or a set of ASCII flat files). These polymorphism data may be stored on the computer's hard drive or may, for example, be stored on a CD-ROM or on one or more other storage devices accessible by the computer. For example, the data may be stored on one or more databases in communication with the computer via a network.

Preferred embodiments of the invention are described in the following examples. Other embodiments within the scope of the claims herein will be apparent to one skilled in the art from consideration of the specification or practice of the invention as disclosed herein. It is intended that the specification, together with the examples, be considered exemplary only, with the scope and spirit of the invention being indicated by the claims which follow the examples.

EXAMPLES

The Examples herein are meant to exemplify the various aspects of carrying out the invention and are not intended to limit the scope of the invention in any way. The Examples do not include detailed descriptions for conventional methods employed, such as in the performance of genomic DNA isolation, PCR and sequencing procedures. Such methods are well-known to those skilled in the art and are described in numerous publications, for example, Sambrook, Fritsch, and Maniatis, "Molecular Cloning: A Laboratory Manual", 2nd Edition, Cold Spring Harbor Laboratory Press, USA, (1989).

EXAMPLE 1

This example illustrates examination of various regions of the SLC26A2 gene for polymorphic sites.

Amplification of Target Regions

The following target regions of the SLC26A2 gene were amplified using PCR primer pairs. The primers used for each region are represented below by providing the nucleotide positions of their initial and final nucleotides, which correspond to positions in Figure 1.

PCR Primer Pairs

Fragment No.	Forward Primer	Reverse Primer	PCR Product
Fragment 1	135711-135734	complement of 136380-136401	691 nt
Fragment 2	136045-136069	complement of 136680-136702	658 nt
Fragment 3	136095-136117	complement of 136680-136702	608 nt
Fragment 4	136312-136336	complement of 136933-136954	643 nt
Fragment 5	136671-136694	complement of 137333-137355	685 nt
Fragment 6	138745-138768	complement of 139392-139413	669 nt
Fragment 7	139104-139125	complement of 139789-139810	707 nt
Fragment 8	139319-139341	complement of 139951-139972	654 nt
Fragment 9	139641-139662	complement of 140203-140224	584 nt
Fragment 10	139917-139937	complement of 140471-140494	578 nt
Fragment 11	140118-140138	complement of 140849-140869	752 nt

These primer pairs were used in PCR reactions containing genomic DNA isolated from immortalized cell lines for each member of the Index Repository. The PCR reactions were carried out under the following conditions:

Reaction volume	= 10 µl
10 x Advantage 2 Polymerase reaction buffer (Clontech)	= 1 µl
100 ng of human genomic DNA	= 1 µl
10 mM dNTP	= 0.4 µl
Advantage 2 Polymerase enzyme mix (Clontech)	= 0.2 µl
Forward Primer (10 µM)	= 0.4 µl
Reverse Primer (10 µM)	= 0.4 µl
Water	= 6.6µl

Amplification profile:
97°C - 2 min. 1 cycle

97°C - 15 sec.
70°C - 45 sec.
72°C - 45 sec. } 10 cycles

97°C - 15 sec.
64°C - 45 sec.
72°C - 45 sec. } 35 cycles

Sequencing of PCR Products

The PCR products were purified using a Whatman/Polyfiltronics 100 µl 384 well unifilter plate essentially according to the manufacturers protocol. The purified DNA was eluted in 50 µl of distilled water. Sequencing reactions were set up using Applied Biosystems Big Dye Terminator chemistry essentially according to the manufacturers protocol. The purified PCR products were sequenced in both directions using the primer sets described previously or those represented below by the nucleotide positions of their initial and final nucleotides, which correspond to positions in Figure 1. Reaction products were purified by isopropanol precipitation, and run on an Applied Biosystems 3700 DNA Analyzer.

Sequencing Primer Pairs

Fragment No.	Forward Primer	Reverse Primer
Fragment 1	135781-135798	complement of 136316-136336
Fragment 2	136101-136122	complement of 136631-136650
Fragment 3	136159-136180	complement of 136638-136657
Fragment 4	136426-136447	complement of 136905-136925
Fragment 5	136717-136736	complement of 137236-137255
Fragment 6	138789-138808	complement of 139315-139335
Fragment 7	139135-139155	complement of 139635-139656
Fragment 8	139389-139408	complement of 139913-139932
Fragment 9	139671-139691	complement of 140126-140146
Fragment 10	139949-139968	complement of 140427-140448
Fragment 11	140211-140230	complement of 140714-140735

Analysis of Sequences for Polymorphic Sites

Sequence information for a minimum of 80 humans was analyzed for the presence of polymorphisms using the Polyphred program (Nickerson et al., *Nucleic Acids Res.* 14:2745-2751, 1997). The presence of a polymorphism was confirmed on both strands. The polymorphisms and their locations in the SLC26A2 gene are listed in Table 3 below.

Table 3. Polymorphic Sites Identified in the SLC26A2 Gene

Polymorphic Site Number	PolyId[a]	Nucleotide Position[b]	Reference Allele	Variant Allele	CDS Variant Position	AA Variant
PS1	3759084	136098	G	A		
PS2	3759090	136195	A	G		
PS3	3759094	139338	T	A	1046	F349Y
PS4[R]	3759102	140013	C	T	1721	T574I
PS5	3759106	140357	A	T	2065	T689S

[a]PolyId is a unique identifier assigned to each PS by Genaissance Pharmaceuticals, Inc.
[b]Positions of polymorphic sites in Figure 1.
[R]Previously reported in the literature

EXAMPLE 2

This example illustrates analysis of the SLC26A2 polymorphisms identified in the Index Repository for human genotypes and haplotypes.

The different genotypes containing these polymorphisms that were observed in the reference population are shown in Table 4 below, with the haplotype pair indicating the combination of haplotypes determined for the individual using the haplotype derivation protocol described below. In Table 4, homozygous positions are indicated by one nucleotide and heterozygous positions are indicated by two nucleotides. Missing nucleotides in any given genotype in Table 4 were inferred based on linkage disequilibrium and/or Mendelian inheritance.

Table 4. Genotypes and Haplotype Pairs Observed for SLC26A2 Gene

Genotype Number	Polymorphic Sites PS1	PS2	PS3	PS4	PS5	HAP	Pair
1	G	A	T	C	A	3	3
2	A	G	T	C	T	2	2
3	A/G	G/A	T	C/T	T/A	2	5
4	G	A	T	C	A/T	3	4
5	G/A	A/G	T/A	C	A/T	3	1
6	G	A	T	C/T	A	3	5
7	G/A	A/G	T	C	A/T	3	2

The haplotype pairs shown in Table 4 were estimated from the unphased genotypes using a computer-implemented extension of Clark's algorithm (Clark, A.G. 1990 *Mol Bio Evol* 7, 111-122) for assigning haplotypes to unrelated individuals in a population sample, as described in U.S. Provisional Application Serial No. 60/198,340 entitled "A Method and System for Determining Haplotypes from a Collection of Polymorphisms" and the corresponding International Application, PCT/US01/12831. In this method, haplotypes are assigned directly from individuals who are homozygous at all sites or heterozygous at no more than one of the variable sites. This list of haplotypes is then used to deconvolute the unphased genotypes in the remaining (multiply heterozygous) individuals. In our analysis, the list of haplotypes was augmented with haplotypes obtained from two families (one three-generation Caucasian family and one two-generation African-American family).

By following this protocol, it was determined that the Index Repository examined herein and, by extension, the general population contains the 5 human SLC26A2 haplotypes shown in Table 5 below.

An SLC26A2 isogene defined by a full-haplotype shown in Table 5 below comprises the regions of the SEQ ID NOS indicated in Table 5, with their corresponding set of polymorphic locations and identities, which are also set forth in Table 5.

Table 5. Haplotypes Identified in the SLC26A2 Gene

Haplotype Number[a]					PS No[b]	PS Position[c]	SEQ ID NO[d]	Region Examined[e]
1	2	3	4	5				
A	A	G	G	G	1	3798 / 30	24 / 25	3411-5055
G	G	A	A	A	2	3895 / 150	24 / 25	3411-5055
A	T	T	T	T	3	7038 / 270	24 / 25	6445-8569
C	C	C	C	T	4	7713 / 390	24 / 25	6445-8569
T	T	A	T	A	5	8057 / 510	24 / 25	6445-8569

[a]Alleles for SLC26A2 haplotypes are presented 5′ to 3′ in each column
[b]PS = polymorphic site;
[c]Position of PS within the indicated SEQ ID NO, with the 1st position number referring to the first SEQ ID NO and the 2nd position number referring to the 2nd SEQ ID NO;
[d]1st SEQ ID NO refers to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol; 2nd SEQ ID NO is a modified version of the 1st SEQ ID NO that comprises the context sequence of each polymorphic site, PS1-PS5, to facilitate electronic searching of the haplotypes;
[e]Region examined represents the nucleotide positions defining the start and stop positions within the 1st SEQ ID NO of the sequenced region.

SEQ ID NO:24 refers to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol. SEQ ID NO:25 is a modified version of SEQ ID NO:24 that shows the context sequence of each of PS1-PS5 in a uniform format to facilitate electronic searching of the SLC26A2 haplotypes. For each polymorphic site, SEQ ID NO:25 contains a block of 60 bases of the nucleotide sequence encompassing the centrally-located polymorphic site at the 30th position, followed by 60 bases of unspecified sequence to represent that each polymorphic site is separated by genomic sequence whose composition is defined elsewhere herein.

Table 6 below shows the percent of chromosomes characterized by a given SLC26A2 haplotype for all unrelated individuals in the Index Repository for which haplotype data was obtained. The percent of these unrelated individuals who have a given SLC26A2 haplotype pair is shown in Table 7. In Tables 6 and 7, the "Total" column shows this frequency data for all of these unrelated individuals, while the other columns show the frequency data for these unrelated individuals categorized according to their self-identified ethnogeographic origin. Abbreviations used in Tables 6 and 7 are AF = African Descent, AS = Asian, CA = Caucasian, HL = Hispanic-Latino, and AM = Native American.

Table 6. Frequency of Observed SLC26A2 Haplotypes In Unrelated Individuals

HAP No.	HAP ID	Total	CA	AF	AS	HL	AM
1	3760324	0.61	2.38	0.0	0.0	0.0	0.0
2	3760318	21.95	19.05	37.5	10.0	19.44	33.33
3	3760317	75.0	76.19	60.0	90.0	75.0	66.67
4	3760323	0.61	0.0	2.5	0.0	0.0	0.0
5	3760322	1.83	2.38	0.0	0.0	5.56	0.0

Table 7. Frequency of Observed SLC26A2 Haplotype Pairs In Unrelated Individuals

HAP1	HAP2	Total	CA	AF	AS	HL	AM
3	3	56.1	57.14	30.0	80.0	61.11	33.33
2	2	4.88	4.76	10.0	0.0	5.56	0.0
2	5	1.22	0.0	0.0	0.0	5.56	0.0
3	4	1.22	0.0	5.0	0.0	0.0	0.0
3	1	1.22	4.76	0.0	0.0	0.0	0.0
3	5	2.44	4.76	0.0	0.0	5.56	0.0
3	2	32.93	28.57	55.0	20.0	22.22	66.67

The size and composition of the Index Repository were chosen to represent the genetic diversity across and within four major population groups comprising the general United States population. For example, as described in Table 1 above, this repository contains approximately equal sample sizes of African-descent, Asian-American, European-American, and Hispanic-Latino population groups. Almost all individuals representing each group had all four grandparents with the same ethnogeographic background. The number of unrelated individuals in the Index Repository

provides a sample size that is sufficient to detect SNPs and haplotypes that occur in the general population with high statistical certainty. For instance, a haplotype that occurs with a frequency of 5% in the general population has a probability higher than 99.9% of being observed in a sample of 80 individuals from the general population. Similarly, a haplotype that occurs with a frequency of 10% in a specific population group has a 99% probability of being observed in a sample of 20 individuals from that population group. In addition, the size and composition of the Index Repository means that the relative frequencies determined therein for the haplotypes and haplotype pairs of the SLC26A2 gene are likely to be similar to the relative frequencies of these SLC26A2 haplotypes and haplotype pairs in the general U.S. population and in the four population groups represented in the Index Repository. The genetic diversity observed for the three Native Americans is presented because it is of scientific interest, but due to the small sample size it lacks statistical significance.

In view of the above, it will be seen that the several advantages of the invention are achieved and other advantageous results attained.

As various changes could be made in the above methods and compositions without departing from the scope of the invention, it is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

All references cited in this specification, including patents and patent applications, are hereby incorporated in their entirety by reference. The discussion of references herein is intended merely to summarize the assertions made by their authors and no admission is made that any reference constitutes prior art. Applicants reserve the right to challenge the accuracy and pertinency of the cited references.

What is Claimed is:

1. A method for haplotyping the solute carrier family 26, member 2 (SLC26A2) gene of an individual, which comprises determining which of the SLC26A2 haplotypes shown in the table immediately below defines one copy of the individual's SLC26A2 gene, wherein each of the SLC26A2 haplotypes comprises a set of polymorphisms whose locations and identities are set forth in the table immediately below:

Haplotype Number[a]					PS No[b]	PS Position[c]
1	2	3	4	5		
A	A	G	G	G	1	3798
G	G	A	A	A	2	3895
A	T	T	T	T	3	7038
C	C	C	C	T	4	7713
T	T	A	T	A	5	8057

[a]Alleles for haplotypes are presented 5′ to 3′ in each column
[b]PS = polymorphic site;
[c]Position of PS within SEQ ID NO:24.

2. The method of claim 1, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS5 on the one copy of the individual's SLC26A2 gene.

3. A method for haplotyping the solute carrier family 26, member 2 (SLC26A2) gene of an individual, which comprises determining which of the SLC26A2 haplotype pairs shown in the table immediately below defines both copies of the individual's SLC26A2 gene, wherein each of the SLC26A2 haplotype pairs consists of first and second haplotypes which comprise first and second sets of polymorphisms whose locations and identities are set forth in the table immediately below:

Haplotype Pairs[a]							PS No[b]	PS Position[c]
3/3	2/2	2/5	3/4	3/1	3/5	3/2		
G/G	A/A	A/G	G/G	G/A	G/G	G/A	1	3798
A/A	G/G	G/A	A/A	A/G	A/A	A/G	2	3895
T/T	T/T	T/T	T/T	T/A	T/T	T/T	3	7038
C/C	C/C	C/T	C/C	C/C	C/T	C/C	4	7713
A/A	T/T	T/A	A/T	A/T	A/A	A/T	5	8057

[a]Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24.

4. The method of claim 3, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS5 on both copies of the individual's SLC26A2 gene.

5. A method for genotyping the solute carrier family 26, member 2 (SLC26A2) gene of an individual, comprising determining for the two copies of the SLC26A2 gene present in the

individual the identity of the nucleotide pair at one or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3 and PS5, wherein the one or more PS have the location and alternative alleles shown in SEQ ID NO:24.

6. The method of claim 5, wherein the determining step comprises:
   (a) isolating from the individual a nucleic acid mixture comprising both copies of the SLC26A2 gene, or a fragment thereof, that are present in the individual;
   (b) amplifying from the nucleic acid mixture a target region containing the selected polymorphic site;
   (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
   (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
   (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.

7. The method of claim 5, which comprises determining for the two copies of the SLC26A2 gene present in the individual the identity of the nucleotide pair at each of PS1-PS5.

8. A method for haplotyping the solute carrier family 26, member 2 (SLC26A2) gene of an individual which comprises determining, for one copy of the SLC26A2 gene present in the individual, the identity of the nucleotide at two or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3 and PS5, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:24.

9. The method of claim 8, further comprising determining the identity of the nucleotide at PS4, which has the location and alternative alleles shown in SEQ ID NO:24.

10. The method of claim 8, wherein the determining step comprises:
   (a) isolating from the individual a nucleic acid sample containing only one of the two copies of the SLC26A2 gene, or a fragment thereof, that is present in the individual;
   (b) amplifying from the nucleic acid sample a target region containing the selected polymorphic site;
   (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
   (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
   (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.

11. A method for predicting a haplotype pair for the solute carrier family 26, member 2 (SLC26A2) gene of an individual comprising:

(a) identifying a SLC26A2 genotype for the individual, wherein the genotype comprises the nucleotide pair at two or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3 and PS5, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:24;

(b) enumerating all possible haplotype pairs which are consistent with the genotype;

(c) comparing the possible haplotype pairs to the haplotype pair data set forth in the table immediately below; and

(d) assigning a haplotype pair to the individual that is consistent with the data

Haplotype Pairs[a]							PS No[b]	PS Position[c]
3/3	2/2	2/5	3/4	3/1	3/5	3/2		
G/G	A/A	A/G	G/G	G/A	G/G	G/A	1	3798
A/A	G/G	G/A	A/A	A/G	A/A	A/G	2	3895
T/T	T/T	T/T	T/T	T/A	T/T	T/T	3	7038
C/C	C/C	C/T	C/C	C/C	C/T	C/C	4	7713
A/A	T/T	T/A	A/T	A/T	A/A	A/T	5	8057

[a]Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24.

12. The method of claim 11, wherein the identified genotype of the individual comprises the nucleotide pair at each of PS1-PS5, which have the location and alternative alleles shown in SEQ ID NO:24.

13. A method for identifying an association between a trait and at least one haplotype or haplotype pair of the solute carrier family 26, member 2 (SLC26A2) gene which comprises comparing the frequency of the haplotype or haplotype pair in a population exhibiting the trait with the frequency of the haplotype or haplotype pair in a reference population, wherein the haplotype is selected from haplotypes 1-5 shown in the table presented immediately below, wherein each of the haplotypes comprises a set of polymorphisms whose locations and identities are set forth in the table immediately below:

Haplotype Number[a]					PS No[b]	PS Position[c]
1	2	3	4	5		
A	A	G	G	G	1	3798
G	G	A	A	A	2	3895
A	T	T	T	T	3	7038
C	C	C	C	T	4	7713
T	T	A	T	A	5	8057

[a]Alleles for haplotypes are presented 5′ to 3′ in each column
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24;

and wherein the haplotype pair is selected from the haplotype pairs shown in the table

immediately below, wherein each of the SLC26A2 haplotype pairs consists of first and second haplotypes which comprise first and second sets of polymorphisms whose locations and identities are set forth in the table immediately below:

Haplotype Pairs[a]							PS No[b]	PS Position[c]
3/3	2/2	2/5	3/4	3/1	3/5	3/2		
G/G	A/A	A/G	G/G	G/A	G/G	G/A	1	3798
A/A	G/G	G/A	A/A	A/G	A/A	A/G	2	3895
T/T	T/T	T/T	T/T	T/A	T/T	T/T	3	7038
C/C	C/C	C/T	C/C	C/C	C/T	C/C	4	7713
A/A	T/T	T/A	A/T	A/T	A/A	A/T	5	8057

[a]Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24;

wherein a higher frequency of the haplotype or haplotype pair in the trait population than in the reference population indicates the trait is associated with the haplotype or haplotype pair.

14. The method of claim 13, wherein the trait is a clinical response to a drug targeting SLC26A2.

15. An isolated genotyping oligonucleotide for detecting a polymorphism in the solute carrier family 26, member 2 (SLC26A2) gene at a polymorphic site (PS) selected from the group consisting of PS1, PS2, PS3 and PS5, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:24.

16. The isolated genotyping oligonucleotide of claim 15, which is an allele-specific oligonucleotide that specifically hybridizes to an allele of the SLC26A2 gene at a region containing the polymorphic site.

17. The allele-specific oligonucleotide of claim 16, which comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:4-7, the complements of SEQ ID NOS:4-7, and SEQ ID NOS:8-15.

18. The isolated genotyping oligonucleotide of claim 15, which is a primer-extension oligonucleotide.

19. The primer-extension oligonucleotide of claim 18,which comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:16-23.

20. A kit for genotyping the solute carrier family 26, member 2 (SLC26A2) gene of an individual, which comprises a set of oligonucleotides designed to genotype each of polymorphic sites (PS) PS1, PS2, PS3 and PS5, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:24.

21. The kit of claim 20, which further comprises oligonucleotides designed to genotype PS4, having the location and alternative alleles shown in SEQ ID NO:24.

22. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:

(a) a first nucleotide sequence which comprises a solute carrier family 26, member 2 (SLC26A2) isogene, wherein the SLC26A2 isogene is selected from the group consisting of isogenes 1- 2 and 4 - 5 shown in the table immediately below and wherein each of the isogenes comprises the regions of the SEQ ID NOS shown in the table immediately below and wherein each of the isogenes 1- 2 and 4 - 5 is further defined by the corresponding set of polymorphisms whose locations and identities are set forth in the table immediately below

Isogene Number[a]					PS No[b]	PS Position[c]	Region Examined[d]
1	2	3	4	5			
A	A	G	G	G	1	3798	3411-5055
G	G	A	A	A	2	3895	3411-5055
A	T	T	T	T	3	7038	6445-8569
C	C	C	C	T	4	7713	6445-8569
T	T	A	T	A	5	8057	6445-8569

[a]Alleles for isogenes are presented 5′ to 3′ in each column
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24;
[d]Region examined represents the nucleotide positions defining the start and stop positions of the sequenced region within SEQ ID NO:24;

(b) a second nucleotide sequence which comprises a fragment of the first nucleotide sequence, wherein the fragment comprises one or more polymorphisms selected from the group consisting of adenine at PS1, guanine at PS2, adenine at PS3 and thymine at PS5, wherein the selected polymorphism has the location set forth in the table immediately above; and

(c) a third nucleotide sequence which is complementary to the first or second nucleotide sequence.

23. The isolated polynucleotide of claim 22, which is a DNA molecule and comprises both the first and third nucleotide sequences and further comprises expression regulatory elements operably linked to the first nucleotide sequence.

24. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 22, wherein the organism expresses a SLC26A2 protein encoded by the first nucleotide sequence.

25. The recombinant nonhuman organism of claim 24, which is a transgenic animal.

26. The isolated polynucleotide of claim 22 which consists of the second nucleotide sequence.

27. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:

(a) a coding sequence for a solute carrier family 26, member 2 (SLC26A2) isogene wherein the coding sequence is defined by a haplotype selected from the group consisting of 1, 2, 4 and 5 shown in the table immediately below and wherein the coding sequence comprises

SEQ ID NO:2 except at each of the polymorphic sites which have the locations and polymorphisms set forth in the table immediately below:

Coding Sequence Haplotype Number[a]			PS No[b]	PS Position[c]
1c	2c, 4c	5c		
A	T	T	3	7038
C	C	T	4	7713
T	T	A	5	8057

[a]Alleles for coding sequence haplotypes are presented 5′ to 3′ in each column; the numerical portion of the coding sequence haplotype number represents the number of the parent full SLC26A2 haplotype;
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:2;

and

(b) a fragment of the coding sequence, wherein the fragment comprises at least one polymorphism selected from the group consisting of adenine at a position corresponding to nucleotide 1046 and thymine at a position corresponding to nucleotide 2065, wherein said positions in the coding sequence and the fragment refer to SEQ ID NO:2.

28. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 27, wherein the organism expresses a solute carrier family 26, member 2 (SLC26A2) protein encoded by the polymorphic variant sequence.

29. The recombinant nonhuman organism of claim 28, which is a transgenic animal.

30. An isolated polypeptide comprising an amino acid sequence which is a polymorphic variant of a reference sequence for the solute carrier family 26, member 2 (SLC26A2) protein or a fragment thereof, wherein the reference sequence comprises SEQ ID NO:3 and the polymorphic variant comprises one or more variant amino acids selected from the group consisting of tyrosine at a position corresponding to amino acid position 349 and serine at a position corresponding to amino acid position 689.

31. An isolated monoclonal antibody specific for and immunoreactive with the isolated polypeptide of claim 30.

32. A method for screening for drugs targeting the isolated polypeptide of claim 30 which comprises contacting the SLC26A2 polymorphic variant with a candidate agent and assaying for binding activity.

33. A computer system for storing and analyzing polymorphism data for the solute carrier family 26, member 2 gene, comprising:

(a) a central processing unit (CPU);

(b) a communication interface;

(c) a display device;

(d) an input device; and

(e) a database containing the polymorphism data;

wherein the polymorphism data comprises the haplotypes set forth in the table immediately below:

Haplotype Number[a]					PS No[b]	PS Position[c]
1	2	3	4	5		
A	A	G	G	G	1	3798
G	G	A	A	A	2	3895
A	T	T	T	T	3	7038
C	C	C	C	T	4	7713
T	T	A	T	A	5	8057

[a]Alleles for haplotypes are presented 5′ to 3′ in each column
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24;

and the haplotype pairs set forth in the table immediately below:

Haplotype Pairs[a]							PS No[b]	PS Position[c]
3/3	2/2	2/5	3/4	3/1	3/5	3/2		
G/G	A/A	A/G	G/G	G/A	G/G	G/A	1	3798
A/A	G/G	G/A	A/A	A/G	A/A	A/G	2	3895
T/T	T/T	T/T	T/T	T/A	T/T	T/T	3	7038
C/C	C/C	C/T	C/C	C/C	C/T	C/C	4	7713
A/A	T/T	T/A	A/T	A/T	A/A	A/T	5	8057

[a]Haplotype pairs are represented as 1st Haplotype/2nd Haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;
[c]Location of PS in SEQ ID NO:24.

34. A genome anthology for the solute carrier family 26, member 2 (SLC26A2) gene which comprises SLC26A2 isogenes defined by any one of haplotypes 1-5 set forth in the table shown below:

Haplotype Number[a]					PS No[b]	PS Position[c]
1	2	3	4	5		
A	A	G	G	G	1	3798
G	G	A	A	A	2	3895
A	T	T	T	T	3	7038
C	C	C	C	T	4	7713
T	T	A	T	A	5	8057

[a]Alleles for haplotypes are presented 5′ to 3′ in each column
[b]PS = polymorphic site;
[c]Position of PS in SEQ ID NO:24.

POLYMORPHISMS IN THE SLC26A2 GENE

GTGTTTATCT GAGAAAGTTT TTATTTCTTT TTCATATTTT ATTTTATTTA
AGACAGAGTC TCTCCCTGTC ACCCAGGCTA AAGTGCAGTG GTGTGATCAT 132400
GGCTTATTGC AGCCTCAACC TCTTAGGCTC AAGCAGTCTC CCTACCTCAG
CCTCCCAAGT AGCTGGGACC ATAGGCACAC ACCACCATGC CTGGCCTCTT 132500
ATCTTTTAAA ATATTTTTTC TAGGTAAAGA ATTCTATGTT AAAAGTTATA
TTTCTGGGGT CAGATGTGGT GGCTCACACC TGTAATCCTA GCACTTTGGG 132600
AGGCCAGGGT GGGAGGATCA CTTGAGGCCA AGAGTTTGGG AACCAGCCTG
GGGAGCTTAG TAGCAAGACC TTGTCTCTAC AAAAAAAATT AAAAAATGAG 132700
CCAGGCTTGG TCGTGCACAC CTGTAATCCT CGCTACTCAG GAGGCTGAGG
TGGGAGAATC ACTTGAGCCC AGGAGTTCGA GGCCATGGTG AGCCATAATT 132800
GCATGTCACT GCACTTCAGC CTGGGCAACA GGGCAAGACC CTATTTCTAT
AAAAGATATA TATTTCTTTG ATACAAAGAT ACCACTCCAT CATGTTCTGG 132900
CCTCCATGAT TTCCAATAAG AAGTCTGCTA TAATTCTTAT TTTTGTTTCT
GTATAAGTAA TATATGTTTT TTCTTTGTCT TCAGGATTTT CTCTTTATCT 133000
TTTGTTTTCA GCAGCTTGAA TATCATATGT GTAGGTGAGG GTTTTTGTTT
TCTTTTTGTT TTTATATTTA CTCTTCTTGG AATTCTCTAA GATTCTTGGA 133100
GATAGATTAT ATATTATATA TATATGTATT TTTTTTTTGA GACAGAGTTT
TGCTTTGTCA CCATGCTGGA GTGCAGTGGC GCGATCTTGG CTCACTGCAG 133200
CCTCCGCCTC CCAGGTTCAA GCAATTCTCC TGCCTCAGCC TCCTGAGTAG
CTGGGACTAC AGGAGCGCAC CACCATGCCC AGCTAATTTT TGTATTTTTT 133300
TAGAAGAGAC GGGGTTTCAC CATGTTGGCC AGGATGGTCT TGATCTCTTG
ACCTTGTGAT CTGCCTGCCT TGGCCTCCGG AAGTGTTGGA ATTACAGGTG 133400
TGAGCCACTG CACCCGGCCT TTTTTTTTTT TTTTTTTTTT GAGAAGGAGT
TTTCCTCTTG CTGTCCAGAC AGGAGTGCAA TGGCGTGATC TCGGCTCACT 133500
GCAACCTCCA GTTCCTGGGT TCCAGCGATT CTCCTGCCTC GGCCTCGCGA
GTAGCTGGGA TTACAGGCGT GTGCCACCAT GCCTGACTAA TTTTATATTT 133600
TTAGTAGAGA CCCTGTTTCA CCATGTTGGT CATGCTGGTC TTGAACTCCT
GGCCTCAGGT TACTTGGATC TATATTTTGA TGTCTCTTAT TAATTTTGGA 133700
AATTTCTTGG CCATTATATC TTGAAATATT TCTTCTGCCC ATTCTCTCTC
TTCTTTTGGT ATTCTAATTA TGCATATGTT TGATACTTTG ATGTGTCCCG 133800
CAGCTCTTGG ATGTTATGTT ACATTGTTGA TATTGGGGTT TTTTTGTTTT
GTTTCGTTAT TTAAACTTTT CTGCGTTTTA GTTTAGCTAA TTTCTGTTTA 133900
TCTGTCTTTT TGAGTTCATC GATTCTTTCT TTGGCTTTCG GGTCTAGTAA
TGATCCTATC GAAGGCATTC ATCATCTCTG TTACTGTGTT TTTCACTTTT 134000
AGCATTCCTT ATTTGATTTT TTCTTACAGT TTCCATTTCT TTGTTTTACT
TTCTATGTGA ATTTTTATCT ATATTCTTCC CAGATAGCAG CCATACTGGG 134100
CAAGATTTGG ATTCCCTTAG TTTTTTCTTA TGTCTCTTTT GCATATCTTA
GGTTTGGAAA AACCTTTTCA CCCAGAAATC ACCAAGAGCT CTCCAAAATA 134200
GGACGTTGGT GCACTCAAGC TAGCTCTAGA AAGCAGGAAA TGACAGGAGA
GTCACAGACT GGTTGCATTG TAGAGCAAAG AGCTGCTGGG TGTATCCATC 134300
AACCTGATTG ATTTTCTCAG GGTCAGTGGT CACCTAAGTC TATTCTTTCT
TTCCTTATCT ACTTCAGAAG ACCATCAAAA TATTTACTCA TGTTAAACCT 134400
AAACAAAATG AATATGGGTT ACTTTTACAT ATCTTGTGTT CTTGAATAAG
TCATAGGTCC TGGATTTTGG CAGTTTGGTA ACGTAAGGTT TATAGAACTT 134500
CACAAGGGCC CATTTTTCTG TCAGTTCCAA AACAAGAAAC TTCCTGAGTC
CTTCATCTAA AATATATTGT CTCGGCTGGG TGCAGTGGCT CGTGCCTGTT 134600
ATCCCAGCAC TTTGGGAGGC TGAGGCGGAT GGATCACCTG AGGTTAGGAC
TTCGAGACCA GCCTGACCAA TATGGTGAAA CCCTGTCTCT ACTAAAAATG 134700
CAAAAATTAG CCAGGCGTGG TGGCATGCAC CTGTAATCCC AGCTACTCAG

FIGURE 1A

```
GAGGCTGAGA TAAGAGAATT GCTTGAACCC AGGAGGTGGA GGTTGTAGTG      134800
AGCTGAGATT GCACCACTGC ACTCCAGCCT GGGCGACAGA GTGAAACTCT
GTCTCAAAAA AAAAAAAAAA AATGTCTCTA TCTGGCCACA GTCACAAATG      134900
TTTGTTCATT TGTTCATTCA TTCATTCAAA TGTTTTGTAA GCCTGCTATC
TCAGCGTTAC TACATTCCAT TCAGATTACA CTGATGAACA AGATGTCTTT      135000
CCTCCAGGAG CTAGAGAGAT TCCTACTTCA CTAATACAAG AGTGTGGTTA
GTACTCTAAT GGAGGTGCAA CATGCTATGG GACACAGAGG GTGTAGTATT      135100
TCATTTGGGC TAGGGGAGAT TGGTTAGTGC TTTCTGGAAA AGGTAGCATT
GTAACTGGGT TTTAAAAAAT TATTAGGATC TTGACAGGCA AAGAGGTGGA      135200
TGGCCATTCG AAGCTAAGTA AACAGCTTAT GTAAAGGCAC TAATTCATGA
AGCATTTGGT AAACAATTTA TGTTCTATTC CTTTGAGAGC CTGGTTCATT      135300
TTCTTCTCTT ACTCCGGTTA TTAGACTTAC TATTTGTTGT TGTCCTTTCT
CTTTTTCTGG CTATTTTTAC CTCCTTTGTT TTCCTATAGT TCCTCATGGT      135400
AGATCTTATG GCATTAGTTT TATAGTCTAG GACACAGAGA TGAAGGATCA
CCTGTATTGC CTCCAAGTGG AAGTGCAGGG CAACATTATT TCTCTATTTA      135500
ACCTGTGTTT CAGTGTGTGT ACTTAGAATA GTAAAGTGAA TCTTGCATGA
ATGTAGGCCC TGCCCACAGG GCAGATGACT CCATACTAGA ACATAGTGGA      135600
ATAGACAAAA ACCTTCTACA GCATGTATGA GACACTTGGC CCATCGACCC
TCTTCATGCC CTTTACATTC AGCACCCTCA TATTGACTTC TCTCTCCTCT      135700
TTCCTACCAA GCAAGGGAGT ACTGTTCAAA GACGCAAATG CATTCTGCCC
TAGTTTCTTT TTATTGCTAA AAACATTTAT CTTTACCCTA CAACCTACTT      135800
TTCTATTTAT TTTCAACATT TAGCAGGTTG TTTAAAAAGG GACCAAAAAA
TAAAACAGGA CCATCTTCCT TGTTTCAGGG ACTGGTAGGC AGGCATTAAG      135900
GTTAAGGTAG GGGTTAAGAC CAGATCCTAT TTTGCAGTCT GCCTGGGAGG
TGAAAAACCT GGGAAGAAGA CCGCTGGTAG CATATGTATG GAAAGGAGAC      136000
AGGCTGCCCT TACATCTTTT CAGGAGGAAA AACTGCCAGG GGGAGCCAGG
CATATATGGA GAAGAATCCT TAATGGTTTA TACTCTTGGG AAGTCCTGTA      136100
                                                    A
CCCAGCCAGT TATTTGCTTT GACTTGGCTG TTTAAGGTCT GGTTCTGGTC
TTTTTTTTTC CCCCTAACCA AGACAAATGA GGCTCAATTA AGGAAAAGGG      136200
                                          G
ACATAAGATA CCTATTCCAA AACTGAATTC CTTTTAACTC TCATGAAATG
ACAAATAGAA TTGTTAGTAT ATGTGAGCAC TGAGAATTAC TTTATTGATG      136300
AACACTGGTA TTTTCTCTGG TGTAGGAAGC TGAACCATCT ATCTCCAGAA
ATGTCTTCAG AAAGTAAAGA GCAACATAAC GTTTCACCCA GAGACTCAGC      136400
     [exon 1: 136351..
TGAAGGAAAT GACAGTTATC CATCTGGGAT CCATCTGGAA CTTCAAAGGG
AATCAAGTAC TGACTTCAAG CAATTTGAGA CCAATGATCA ATGCAGACCT      136500
TATCATAGGA TCCTTATTGA GCGTCAAGAG AAATCAGATA CAAACTTCAA
GGAGTTTGTT ATTAAAAAGC TGCAGAAGAA TTGCCAGTGC AGTCCAGCCA      136600
AAGCCAAAAA TATGATTTTA GGTTTCCTTC CTGTTTTGCA GTGGCTCCCA
AAATACGACC TAAAGAAAAA CATTTTAGGG GATGTGATGT CAGGCTTGAT      136700
TGTGGGCATA TTATTGGTGC CCCAGTCCAT TGCTTATTCC CTGCTGGCTG
GCCAAGAACC TGTCTATGGT CTGTACACAT CTTTTTTTGC CAGCATCATT      136800
TATTTTCTCT TGGGTACCTC CCGTCACATC TCTGTGGGCA TTTTTGGAGT
ACTGTGCCTT ATGATTGGTG AGACAGTTGA CCGAGAACTA CAGAAAGCTG      136900
GCTATGACAA TGCCCATAGT GCTCCTTCCT TAGGAATGGT TTCAAATGGG
AGCACATTAT TAAATCATAC ATCAGACAGG ATATGTGACA AAAGTTGCTA      137000
TGCAATTATG GTTGGCAGCA CTGTAACCTT TATAGCTGGA GTTTATCAGG
              ..137049]
```

FIGURE 1B

```
TAAGCAGCAA TGAAACAATT GGTTATTTCT AGAAAAGTAA TCTAGTACAT      137100
GAAATCTCAT ATCTCTAAGG GATCTGAGGA ATCACAATAA TTAAAGGTAT
CATTTATTGA GAGTTCAGGA TATATGAAGG GTAGAGGCAA AATTCAAACC      137200
CTAACCTGAC TCCACAGGTA ATATAAGGCT GGTTCACTGG ACCTCCACCA
CCCAGTACAA CTCCTTAATT TTACATGTCA GAAAATCTTG GCTTTGCTTG      137300
AGATTATTTG TGGCTGGTTA TTGGCAGAGT CAGCATTAGC AGTTAGGCAA
GTGGGTAACA GAATGGAGTT GAGAGTGCAG GAGTTTCTCA CTTTTTTTTT      137400
TTTTCTGGAG ACAGGGTCTC ACTCTGTCAC GCTGGAGTGC AGTGGCACTA
TCTTAGTTCA CTGCAACGTC CGCCTCCCTG GCTCAAGCAG TCCTCCTACC      137500
TCAACCTCCT GAGTAGCTAG GACTACAGGC ACATGCTACC ACACCTGGCT
AATTTTATTT TATTTTATTT TATTTTATTT TTTATTTTTA TTTTTTGTAG      137600
AGACAGGGTT TTGCCACGTT GCCCAGGCTG GTTTCAAACT CCTGAGCTCA
AGCAATCCTC CCGTCTTGGC CTCCCAAAGT GCTGGGATTA TAGCCATGAG      137700
CCACCACACC CAGCCTCAAA TTCTAAATGT CTCTTACCTT CCATTAAAAT
TGCTGATCTA TTGAGCAACT CTTACTAAAG GTAGTGGTTG TCTTGGATTG      137800
TTGGGGAGGG AGGGAAAAAG TTGGGGACCA CAGTTTCATA TTATCAGCCA
GGAGAAAGGA TAAGAAATCA AATTCTTGAG TCTCCCATAG AATCCACTAA      137900
TCTGTCATTA TCATCATGCC CCTGGCTTTT GGCATCCAGG AGTCAGTGCC
AGGATTAAAC CTTCTCTAAT GCAGGCATTT CAAACCAACA AGGGAAGGGG      138000
AAGAGTAGCT CACTTTAGTT GGTGCTCAGA TGAGTGGGGA GGGAGAGTGA
AGATGGTGTG AAGATGAGCT GTCTACTCAT ATATAATGGT AAATAATAAG      138100
TCTACTTACT TATTTATTAT TTATTCATTT ATTTATAAAG AGACAGGGTC
TCTCTATGAC CAAACTCCTG GGCTCAAGTG ATCCTCCTAA TATTGCCTCC      138200
CCAAATGCTG GGATTACAGG CATGAGCCAT CACGCCCAAC CAACTTTTGC
CTTTTTGTTA GTATGTCCCA CCAAGAAGGA AGAAGGCATA ACAATTCTGA      138300
AAACTTATTA GACAGAGGAA AATATAAAGA AGTAAAAATG CAGAATTTTT
ATTAATATGG GAGACAGTGT GGCATAAGTA CATATATACT GCATGAGAAT      138400
GGTTTCTTAG TATGAGGTTA AAGATAAGTC TACAATAATT TTTAAAGTGT
GATTCTACTT TGATGTAAAT CTAATTTTTT GTTTTACCAA TTAAAACTTC      138500
ACTTGTACAC TTGCTCTTAG CCAAGAGGCT GAGAAGCCGT AAGACTTCAC
TTTTACAGTA GTGATTTGTA ATTTAAGGAA AATACTTGGT TTCTTAACTA      138600
GAATAATTTT TTCCAATTTG AAGTTTTCTT GTGGATCCTT GAGAATGTTT
TTCTTTTAAA AGAGGTCTGT TCTTTGTGAT GGGAAGAATG AAAAAAAAAA      138700
GAGGTATGAA CCTTATTCAA GTTTAAGAAA CGTATGAAAA GAAAGAAATC
CAAAGTTCCT GTCTCACCTG GGTTAATAAG TAACAGTGTG ACCTTGGGCA      138800
AGTTGCTTAG CCCTTTAAAC ATAATTTTCA TCTTTGTAAA ATGAGAAGAT
TGATATATGA TTGTGTTTAT TCTAGCTCTG ACATTCTGTG ATGCTCTGAT      138900
GATATGTCTC CATGCAAGAA ATGTCAGGAT AATATAAAAT TTAGAAGTTC
TTTTCCATTT ATATTTAACA CTTCTATATC CTTCCTTCCA GGTAGCGATG      139000
     [exon 2: 138992..
GGCTTCTTTC AAGTGGGTTT TGTTTCTGTC TACCTCTCAG ATGCCTTGCT
GAGTGGATTT GTCACTGGTG CCTCCTTCAC TATTCTTACA TCTCAGGCCA      139100
AGTATCTTCT TGGGCTCAAC CTTCCTCGGA CTAATGGTGT GGGCTCACTC
ATCACTACCT GGATACATGT CTTCAGAAAC ATCCATAAGA CCAATCTCTG      139200
TGATCTTATC ACCAGCCTTT TGTGCCTTTT GGTTCTTTTG CCAACCAAAG
AACTCAATGA ACACTTCAAA TCCAAGCTTA AGGCACCGAT TCCTATTGAA      139300
CTTGTTGTTG TTGTAGCAGC CACATTAGCC TCTCATTTTG GAAAACTACA
                                        A
TGAAAATTAT AATTCTAGTA TTGCTGGACA TATTCCCACT GGGTTTATGC      139400
CACCCAAAGT ACCAGAATGG AACCTAATTC CTAGTGTGGC TGTAGATGCA
ATAGCTATTT CCATCATTGG TTTTGCTATC ACTGTATCAC TTTCTGAGAT      139500
GTTTGCCAAG AAACATGGTT ACACAGTCAA AGCAAACCAG GAAATGTATG
```

FIGURE 1C

```
CCATTGGCTT TTGTAATATC ATCCCTTCCT TCTTCCACTG TTTTACTACT      139600
AGTGCAGCTC TTGCAAAGAC ATTGGTTAAA GAATCAACAG GCTGCCATAC
TCAGCTTTCT GGTGTGGTAA CAGCCCTGGT TCTTTTGTTG GTCCTCCTAG      139700
TAATAGCTCC TTTGTTCTAT TCCCTTCAAA AAAGTGTCCT TGGTGTGATC
ACAATTGTAA ATCTACGGGG AGCCCTTCGT AAATTTAGGG ATCTTCCCAA      139800
AATGTGGAGT ATTAGTAGAA TGGATACAGT TATCTGGTTT GTTACTATGC
TGTCCTCTGC ACTGCTAAGT ACTGAAATAG GCCTACTTGT TGGGGTTTGT      139900
TTTTCTATAT TTTGTGTCAT CCTCCGCACT CAGAAGCCAA AGAGTTCACT
GCTTGGCTTG GTGGAAGAGT CTGAGGTCTT TGAATCTGTG TCTGCTTACA      140000
AGAACCTTCA GACTAAGCCA GGCATCAAGA TTTTCCGCTT TGTAGCCCCT
             T
CTCTACTACA TAAACAAAGA ATGCTTTAAA TCTGCTTTAT ACAAACAAAC      140100
TGTCAACCCA ATCTTAATAA AGGTGGCTTG GAAGAAGGCA GCAAAGAGAA
AGATCAAAGA AAAAGTAGTG ACTCTTGGTG GAATCCAGGA TGAAATGTCA      140200
GTGCAACTTT CCCATGATCC CTTGGAGCTG CATACTATAG TGATTGACTG
CAGTGCAATT CAATTTTTAG ATACAGCAGG GATCCACACA CTGAAAGAAG      140300
TTCGCAGAGA TTATGAAGCC ATTGGAATCC AGGTTCTGCT GGCTCAGTGC
AATCCCACTG TGAGGGATTC CCTAACCAAC GGAGAATATT GCAAAAAGGA      140400
      T
AGAAGAAAAC CTTCTCTTCT ATAGTGTGTA TGAAGCGATG GCTTTTGCAG
AAGTATCTAA AAATCAGAAA GGAGTATGTG TTCCCAATGG TCTGAGTCTT      140500
AGTAGTGATT AATTGAGAAG GTAGATAGAA GAATGTCTAG CCAATAGGTT
                  ..140512]
AAAATTTCAA GTGTCCAACA TTTCCCAGTT CCACAGTGGG AAATTTTGCA      140600
CACTTGAAAT TTTAACCAAG TGGCTAGATA TTATTCCTCC TTTGAAGCTA
ATGGCATTTG TATATACACA CTGCAGCAGA GCTTGTAGCT GGACAGAGTC      140700
AAAAAGAAGA AAATACGGTT TCAGGCTTTC TTGCAGATAT GAAGTATTCT
TGGAATGCAA TAAGTATGTA TTGAACTGTA CTGTAAAGTA GCTCCAAAAC      140800
TTAATTACTC TCCTGTTTTA GGGGTTATAC ATTTGGACTG TGCATTCTCC
AAGAGATGAA GCGGTGAAGT TGGGATTTAC ATTGGAAGTG CTGTAGACTT      140900
CTTTATGTGG CTCAGTGGAG AGAGGGAAAG AATGTTGCAC CTGCTCTAGT
ACCATAGGTC AAGAGGCTTC TGGATCACAA AGTCATAACT AGACAGGTTT      141000
GTTCTTGTAG TTTTCTATCC CCAGTCTTTG CTCCCCAGAT GGCAGTAGTT
TTTAGTAGGA AAGTGCCATT CCTGTCCTTA AGGCACAGTC TCATCAGAAG      141100
TCTAATACCT GGGCAGGTTT ATAACATCCT GAGAGCCAGC CTGACATTAG
ACAGAATACC CTTTGTAATA CATTGGAAAT TTTTACTCAT GCCTTTTTGT      141200
TTAGGATAAA TAGGTAAGCA CAAAGAGCTC TTCAAAATCA GAAAAAACAA
TAGGAGTCCT TCCTTGTCTT TTCTGTGATC TCTGTCCTTG TTTCTGAGAC      141300
TTTCTCTACC ATTAAGCTCT ATTTTAGCTT TCAGTTATTC TAGTTTGTTT
CCCATGGAAT CTGTCCTAAA CTGGTGTTTT TGTCAGTGAC AGTCTTGCCA      141400
GTCAGCAATT TCTAACAGCA TTTTAAATGA GTTTGATGTA CAGTAAATAT
TGATGACAAT GACAGCTTTT AACTCTTCAA GTCACCTAAA GCTATTATGC      141500
AGGAGGATTT AGAAGTCACA TTCATAAAAC CCAAGGGCTA TGGGTGTATT
ATTCATGATA GCTGGCCCAC AGGTCATGAA TTGAGGAGGA ATTTGCTTTC      141600
AAAAAGCAAG AATGTCCAAC ACTGAAAGTT TATAGTTTTA TATTTGGACC
TTGAAAGGTA AGAAAAAACC AGGTTCTCCA AAGTTAGGAA TAGGGAACTA      141700
ATTTATGAAA CAGCCATCTT AAAAAAAAAA AAAGTAAACT GCAAAAGTAC
AAAATCATTT TTCAATCTGT TCCCAGTTTC TAAACAATTT TAAATATTTA      141800
TGAGAAGCAA ACCCTATGTG TAGGGCATCT GTTGGAGTGG GATGCTTTTA
GACATATATT AAGTATGTAC ATGTTTAATA TGTATATTTA AAATGCATAT      141900
ATATTTTATT ATATCTATAT TATCCTATAT AGATATATGT AACTTAGCTT
```

FIGURE 1D

```
TATTGTTAGC TCCATAAGCT GCCAGTGTTG CTTTTCTGTT GGTAGAGCTC      142000
TCCCATTTGG TGACATGGAA AATACCTTTC CATTATCACA ACAAAGCAGT
TGCTCAGTAG AAAGTCTAGA TTTCTGTCTT ATAGGTGATT TCTGTCTTAT      142100
AGGTGATTAT AATCAAGTGT AGGCTTCCTG AATTTTGACA TCCTTTTAGA
ACTTGGGTCT GGAATTCCAG AAATGTTAAT TGCTGCTTGT ATTTGTTCTT      142200
GTTTGTTTTT TAGCCAGTAT TTGCCCTTTC TATCCAGCCT TATGAATAAT
AGCAGTAAAA TCACAGTATC TTGGTCAGTC TTTATTTTTT TCCTTTTTTC      142300
TTTTTTAAGA GACAGTCATC CAGGCCAGAG TGCAGTTTGA TGATAGCTTA
CTGAAGCTTC CCACTCCTGG GCTCAAGTTA TCCTTCCATT TTGGCCTCCT      142400
GAGTAGCTAG ACCATAGGTA TGCATCACCA CACCCTGCTA ATTTTTTAAA
TTTTTTTCTA GAGAGAGGGT CTCACTGTGT TGCCCAGGCT GGTCTCAAAC      142500
TCCAGGCTCA AGCAATCCTT CAGCCTCAGC CTCCCAGAGT GTTGGGATTA
CAGGCGTGAG CCACTGCACT TGGCCAAGTT ATTTATTTTT AATCTCTCTT      142600
GCCCTTCTCC CAAGGCAGGC TTAAGTTGAG ACTATTATAG GTGTCTAATA
ACCTGTGACA GAGTAATGAG TACATGCTTA AGATGTTATA ATTAGCCAAC      142700
ACCAACACAG CAAAAAATAT AATTCCAGCC AAAGATTCTG GAAAATCCCT
CAGAAGGAGG GATAACAGGA TTTGACCTTT ACCAGCGATT TCTGTCCATA      142800
TGTGGATGTA AACAGTTCTG GAACGTTATG CATGCAGTTA GCGAATCCTT
GAATTATGTT CTGGTTTGTA CTTGTCCCAT CCATCCAAAC AAGAGATTCT      142900
GCTTTTGGTA GCCATCTGTA GAAACATTTA AGATGTCACT AGAATTTACA
TTTCATCCTC TCTACTTGGG TTGAGGTTGC CTATACTTGC ATATTGTTAA      143000
AATGTTTTGG TTGCTGATAT TCAGAGGAAT GAAACCTGGA ACCAAAGCCT
AATTTGCCGA TAAAAAAACT GTTTTCGGCC AGGTGCAGTG GCTCATGCCT      143100
GTAATCCCAG CACGTTGGGA GGCCGAGGCG GGTGGATCAC CTGAAGTCAG
GAGTTCGAGA CCATCCTGGC TAACACTGTG AAACCCCGTC TCTACTAAAA      143200
ATACAAAAAA TTAGCGGGGC ATGGTGGCAC GCGCCTGTAG TCCCAGCTAC
TCAGGAGGCT GAGGCAGGAG AATTACTTGA ACCCGGGAGG CGGAGGTTGC      143300
AGTAAGCAGA GATTGTGCCA CTGGACTCCA GCCTGGGTGA CAGAGCGAGA
CTCCGTCTCA AACAAACAAA CAAAAAACTG TTTTCATTTG CTCTCTTGAC      143400
CAAAGGATAG GACTTTAGTT CTTTAAGCAT TATTTTAAAC ACTATATTGA
TACAAAAATA TCTTGCTTAC TCTAAACTTT AGAGTCTAAA TGAAGCTTTT      143500
TCTCAGTACA AGATTCTGAG TATCATAAAA TGGTTATTTA ATTGAAACGT
AGTGTGGTAT ACTCTTGATG GTTAGAACTC TTACAGCCTT ATTTATTTTT      143600
AAGTTTGTTA CAGCCAAAGG GTTGGAGTGT GCCAGTGCAC AGGTAGACTA
AGGAAAACAT TATAGAGGAG TGAAGAGAAC AGACCATTGA AAAGACTATT      143700
ATCTGACCAG CGGAGGCAGA AAAGAGAGGA ACCCAGTTGA ATAGGATCCA
ATCCCTGGTT AGCCTCTACA CAATAATAGG GAGACAAGGA TTAGGAGCCA      143800
TACCTCCCAG AGCAAGGTAT CTTTCTAGAG CAAATTTCTC TTTCTAGAAG
GGGAGGGTCA CAGGGTCACA GATTCACCAA AGCTGAAAGG GCTGAGGAGC      143900
TCATGGTAGC CTGGGTTGAC CTACTCTGGA GCACGGTGTC TTCCTTCTAA
ACTGAGTGAC TGTAGTACTA TCTGTGCCTC TGATGGTAAT AAAACTGACA      144000
AGATGTCTAA TTTTTTTTTA AGTAGGACCA AAGGAAAACA AGATTTAGAT
AGTCTGACTT TGCTTTTGAA CAACAGACAT TGCAAGTCAA AATTGTTGTC      144100
AAATTTACAT ATGGTAAATG ATGAACTTTA AAAATGTGTC CAGGTGTTAG
ATGAGTTCAT TAGACTCTTT TAATGCTAAT GGCTAGTACG TTTAAACAAA      144200
ACAGCAGTTC TCTGCTGCAA TATTCCCATT GACCACTTAA ATGACCATAA
GTGGTCATTT AAGAACATGT TAGGGTTAGC CCTGATCTGA ATATAAAAGT      144300
GAGAAAAGGG CTACAGTGCA TTTCTTGGTA ACTTAAACTG AGTCTTGAAG
TTATAATGAT CCATTCGAGT TCTGTGATCC TTATTGTTCT TAATTGTGTT      144400
TCTCTACGTA TTGTTACAGA TGAGCCATAC GTTTCTTTGT ATCAATGTAG
ACATGACTTC AGATACCTCT GAGGACCTAC CCAGCAGTCT AGGACCCTGG      144500
GCCAAGTGCT GG                                               144512
```

FIGURE 1E

POLYMORPHISMS IN THE CODING SEQUENCE OF SLC26A2

```
ATGTCTTCAG AAAGTAAAGA GCAACATAAC GTTTCACCCA GAGACTCAGC
TGAAGGAAAT GACAGTTATC CATCTGGGAT CCATCTGGAA CTTCAAAGGG      100
AATCAAGTAC TGACTTCAAG CAATTTGAGA CCAATGATCA ATGCAGACCT
TATCATAGGA TCCTTATTGA GCGTCAAGAG AAATCAGATA CAAACTTCAA      200
GGAGTTTGTT ATTAAAAAGC TGCAGAAGAA TTGCCAGTGC AGTCCAGCCA
AAGCCAAAAA TATGATTTTA GGTTTCCTTC CTGTTTTGCA GTGGCTCCCA      300
AAATACGACC TAAAGAAAAA CATTTTAGGG GATGTGATGT CAGGCTTGAT
TGTGGGCATA TTATTGGTGC CCCAGTCCAT TGCTTATTCC CTGCTGGCTG      400
GCCAAGAACC TGTCTATGGT CTGTACACAT CTTTTTTTGC CAGCATCATT
TATTTTCTCT TGGGTACCTC CCGTCACATC TCTGTGGGCA TTTTTGGAGT      500
ACTGTGCCTT ATGATTGGTG AGACAGTTGA CCGAGAACTA CAGAAAGCTG
GCTATGACAA TGCCCATAGT GCTCCTTCCT TAGGAATGGT TTCAAATGGG      600
AGCACATTAT TAAATCATAC ATCAGACAGG ATATGTGACA AAAGTTGCTA
TGCAATTATG GTTGGCAGCA CTGTAACCTT TATAGCTGGA GTTTATCAGG      700
TAGCGATGGG CTTCTTTCAA GTGGGTTTTG TTTCTGTCTA CCTCTCAGAT
GCCTTGCTGA GTGGATTTGT CACTGGTGCC TCCTTCACTA TTCTTACATC      800
TCAGGCCAAG TATCTTCTTG GGCTCAACCT TCCTCGGACT AATGGTGTGG
GCTCACTCAT CACTACCTGG ATACATGTCT TCAGAAACAT CCATAAGACC      900
AATCTCTGTG ATCTTATCAC CAGCCTTTTG TGCCTTTTGG TTCTTTTGCC
AACCAAAGAA CTCAATGAAC ACTTCAAATC CAAGCTTAAG GCACCGATTC     1000
CTATTGAACT TGTTGTTGTT GTAGCAGCCA CATTAGCCTC TCATTTTGGA
                                                    A
AAACTACATG AAAATTATAA TTCTAGTATT GCTGGACATA TTCCCACTGG     1100
GTTTATGCCA CCCAAAGTAC CAGAATGGAA CCTAATTCCT AGTGTGGCTG
TAGATGCAAT AGCTATTTCC ATCATTGGTT TTGCTATCAC TGTATCACTT     1200
TCTGAGATGT TTGCCAAGAA ACATGGTTAC ACAGTCAAAG CAAACCAGGA
AATGTATGCC ATTGGCTTTT GTAATATCAT CCCTTCCTTC TTCCACTGTT     1300
TTACTACTAG TGCAGCTCTT GCAAAGACAT TGGTTAAAGA ATCAACAGGC
TGCCATACTC AGCTTTCTGG TGTGGTAACA GCCCTGGTTC TTTTGTTGGT     1400
CCTCCTAGTA ATAGCTCCTT TGTTCTATTC CCTTCAAAAA AGTGTCCTTG
GTGTGATCAC AATTGTAAAT CTACGGGGAG CCCTTCGTAA ATTTAGGGAT     1500
CTTCCCAAAA TGTGGAGTAT TAGTAGAATG GATACAGTTA TCTGGTTTGT
TACTATGCTG TCCTCTGCAC TGCTAAGTAC TGAAATAGGC CTACTTGTTG     1600
GGGTTTGTTT TTCTATATTT TGTGTCATCC TCCGCACTCA GAAGCCAAAG
AGTTCACTGC TTGGCTTGGT GGAAGAGTCT GAGGTCTTTG AATCTGTGTC     1700
TGCTTACAAG AACCTTCAGA CTAAGCCAGG CATCAAGATT TTCCGCTTTG
                    T
TAGCCCCTCT CTACTACATA AACAAAGAAT GCTTTAAATC TGCTTTATAC     1800
AAACAAACTG TCAACCCAAT CTTAATAAAG GTGGCTTGGA AGAAGGCAGC
AAAGAGAAAG ATCAAAGAAA AAGTAGTGAC TCTTGGTGGA ATCCAGGATG     1900
AAATGTCAGT GCAACTTTCC CATGATCCCT TGGAGCTGCA TACTATAGTG
ATTGACTGCA GTGCAATTCA ATTTTTAGAT ACAGCAGGGA TCCACACACT     2000
GAAAGAAGTT CGCAGAGATT ATGAAGCCAT TGGAATCCAG GTTCTGCTGG
CTCAGTGCAA TCCCACTGTG AGGGATTCCC TAACCAACGG AGAATATTGC     2100
          T
AAAAAGGAAG AAGAAAACCT TCTCTTCTAT AGTGTGTATG AAGCGATGGC
TTTTGCAGAA GTATCTAAAA ATCAGAAAGG AGTATGTGTT CCCAATGGTC     2200
TGAGTCTTAG TAGTGATTAA                                      2220
```

FIGURE 2

ISOFORMS OF THE SLC26A2 PROTEIN

```
MSSESKEQHN VSPRDSAEGN DSYPSGIHLE LQRESSTDFK QFETNDQCRP
YHRILIERQE KSDTNFKEFV IKKLQKNCQC SPAKAKNMIL GFLPVLQWLP      100
KYDLKKNILG DVMSGLIVGI LLVPQSIAYS LLAGQEPVYG LYTSFFASII
YFLLGTSRHI SVGIFGVLCL MIGETVDREL QKAGYDNAHS APSLGMVSNG      200
STLLNHTSDR ICDKSCYAIM VGSTVTFIAG VYQVAMGFFQ VGFVSVYLSD
ALLSGFVTGA SFTILTSQAK YLLGLNLPRT NGVGSLITTW IHVFRNIHKT      300
NLCDLITSLL CLLVLLPTKE LNEHFKSKLK APIPIELVVV VAATLASHFG
                                                        Y
KLHENYNSSI AGHIPTGFMP PKVPEWNLIP SVAVDAIAIS IIGFAITVSL      400
SEMFAKKHGY TVKANQEMYA IGFCNIIPSF FHCFTTSAAL AKTLVKESTG
CHTQLSGVVT ALVLLLVLLV IAPLFYSLQK SVLGVITIVN LRGALRKFRD      500
LPKMWSISRM DTVIWFVTML SSALLSTEIG LLVGVCFSIF CVILRTQKPK
SSLLGLVEES EVFESVSAYK NLQTKPGIKI FRFVAPLYYI NKECFKSALY      600
                        I
KQTVNPILIK VAWKKAAKRK IKEKVVTLGG IQDEMSVQLS HDPLELHTIV
IDCSAIQFLD TAGIHTLKEV RRDYEAIGIQ VLLAQCNPTV RDSLTNGEYC      700
                                          S
KKEEENLLFY SVYEAMAFAE VSKNQKGVCV PNGLSLSSD                  739
```

FIGURE 3

SEQUENCE LISTING

<110> Genaissance Pharmaceuticals, Inc.
Kliem, Stefanie E.
Koshy, Beena
Tanguay, Debra A.

<120> HAPLOTYPES OF THE SLC26A2 GENE

<130> SLC26A2 MWH0849-PCT

<140> TBA
<141> 2001-06-22

<150> 60/213,284
<151> 2000-06-22

<160> 25

<170> PatentIn Ver. 2.1

<210> 1
<211> 12212
<212> DNA
<213> Homo sapiens

<400> 1
gtgtttatct gagaaagttt ttattctttt ttcatatttt attttattta agacagagtc 60
tctccctgtc acccaggcta aagtgcagtg gtgtgatcat ggcttattgc agcctcaacc 120
tcttaggctc aagcagtctc cctacctcag cctcccaagt agctgggacc ataggcacac 180
accaccatgc ctggcctctt atcttttaaa atatttttc taggtaaaga attctatgtt 240
aaaagttata tttctggggt cagatgtggt ggctcacacc tgtaatccta gcactttggg 300
aggccagggt gggaggatca cttgaggcca agagtttggg aaccagcctg gggagcttag 360
tagcaagacc ttgtctctac aaaaaaaatt aaaaaatgag ccaggcttgg tcgtgcacac 420
ctgtaatcct cgctactcag gaggctgagg tgggagaatc acttgagccc aggagttcga 480
ggccatggtg agccataatt gcatgtcact gcacttcagc ctgggcaaca gggcaagacc 540
ctatttctat aaaagatata tatttctttg atacaaagat accactccat catgttctgg 600
cctccatgat ttccaataag aagtctgcta taattcttat ttttgtttct gtataagtaa 660
tatatgtttt ttctttgtct tcaggatttt ctctttatct tttgttttca gcagcttgaa 720
tatcatatgt gtaggtgagg gtttttgttt tctttttgtt tttatattta ctcttcttgg 780
aattctctaa gattcttgga gatagattat atattatata tatatgtatt tttttttga 840
gacagagttt tgctttgtca ccatgctgga gtgcagtggc gcgatcttgg ctcactgcag 900
cctccgcctc ccaggttcaa gcaattctcc tgcctcagcc tcctgagtag ctgggactac 960
aggagcgcac caccatgccc agctaatttt tgtatttttt tagaagagac ggggtttcac 1020
catgttggcc aggatggtct tgatctcttg accttgtgat ctgcctgcct tggcctccgg 1080
aagtgttgga attacaggtg tgagccactg cacccggcct tttttttttt ttttttttt 1140
gagaaggagt tttcctcttg ctgtccagac aggagtgcaa tggcgtgatc tcggctcact 1200
gcaacctcca gttcctgggt tccagcgatt ctcctgcctc ggcctcgcga gtagctggga 1260

ttacaggcgt gtgccaccat gcctgactaa ttttatattt ttagtagaga ccctgtttca 1320
ccatgttggt catgctggtc ttgaactcct ggcctcaggt tacttggatc tatattttga 1380
tgtctcttat taattttgga aatttcttgg ccattatatc ttgaaatatt tcttctgccc 1440
attctctctc ttcttttggt attctaatta tgcatatgtt tgatactttg atgtgtcccg 1500
cagctcttgg atgttatgtt acattgttga tattggggtt tttttgtttt gtttcgttat 1560
ttaaactttt ctgcgtttta gtttagctaa tttctgttta tctgtctttt tgagttcatc 1620
gattctttct ttggctttcg ggtctagtaa tgatcctatc gaaggcattc atcatctctg 1680
ttactgtgtt tttcactttt agcattcctt atttgatttt ttcttacagt ttccatttct 1740
ttgttttact ttctatgtga atttttatct atattcttcc cagatagcag ccatactggg 1800
caagatttgg attcccttag ttttttctta tgtctctttt gcatatctta ggtttggaaa 1860
aacctttca cccagaaatc accaagagct ctccaaaata ggacgttggt gcactcaagc 1920
tagctctaga aagcaggaaa tgacaggaga gtcacagact ggttgcattg tagagcaaag 1980
agctgctggg tgtatccatc aacctgattg attttctcag ggtcagtggt cacctaagtc 2040
tattctttct ttccttatct acttcagaag accatcaaaa tatttactca tgttaaacct 2100
aaacaaaatg aatatgggtt acttttacat atcttgtgtt cttgaataag tcataggtcc 2160
tggattttgg cagtttggta acgtaaggtt tatagaactt cacaagggcc catttttctg 2220
tcagttccaa aacaagaaac ttcctgagtc cttcatctaa aatatattgt ctcggctggg 2280
tgcagtggct cgtgcctgtt atcccagcac tttgggaggc tgaggcggat ggatcacctg 2340
aggttaggac ttcgagacca gcctgaccaa tatggtgaaa ccctgtctct actaaaaatg 2400
caaaaattag ccaggcgtgg tggcatgcac ctgtaatccc agctactcag gaggctgaga 2460
taagagaatt gcttgaaccc aggaggtgga ggttgtagtg agctgagatt gcaccactgc 2520
actccagcct gggcgacaga gtgaaactct gtctcaaaaa aaaaaaaaaa aatgtctcta 2580
tctggccaca gtcacaaatg tttgttcatt tgttcattca ttcattcaaa tgttttgtaa 2640
gcctgctatc tcagcgttac tacattccat tcagattaca ctgatgaaca agatgtcttt 2700
cctccaggag ctagagagat tcctacttca ctaatacaag agtgtggtta gtactctaat 2760
ggaggtgcaa catgctatgg gacacagagg gtgtagtatt tcatttgggc taggggagat 2820
tggttagtgc tttctggaaa aggtagcatt gtaactgggt tttaaaaaat tattaggatc 2880
ttgacaggca aagaggtgga tggccattcg aagctaagta aacagcttat gtaaaggcac 2940
taattcatga agcatttggt aaacaattta tgttctattc ctttgagagc ctggttcatt 3000
ttcttctctt actccggtta ttagacttac tatttgttgt tgtcctttct ctttttctgg 3060
ctatttttac ctcctttgtt ttcctatagt tcctcatggt agatcttatg gcattagttt 3120
tatagtctag gacacagaga tgaaggatca cctgtattgc ctccaagtgg aagtgcaggg 3180
caacattatt tctctattta acctgtgttt cagtgtgtgt acttagaata gtaaagtgaa 3240
tcttgcatga atgtaggccc tgcccacagg gcagatgact ccatactaga acatagtgga 3300
atagacaaaa accttctaca gcatgtatga gacacttggc ccatcgaccc tcttcatgcc 3360
ctttacattc agcaccctca tattgacttc tctctcctct ttcctaccaa gcaagggagt 3420
actgttcaaa gacgcaaatg cattctgccc tagtttcttt ttattgctaa aaacatttat 3480
ctttacccta caacctactt ttctatttat tttcaacatt tagcaggttg tttaaaaagg 3540
gaccaaaaaa taaaacagga ccatcttcct tgtttcaggg actggtaggc aggcattaag 3600
gttaaggtag gggttaagac cagatcctat tttgcagtct gcctgggagg tgaaaaacct 3660
gggaagaaga ccgctggtag catatgtatg gaaaggagac aggctgccct tacatctttt 3720
caggaggaaa aactgccagg gggagccagg catatatgga gaagaatcct taatggttta 3780
tactcttggg aagtcctrta cccagccagt tatttgcttt gacttggctg tttaaggtct 3840
ggttctggtc ttttttttttc cccctaacca agacaaatga ggctcaatta aggaraaggg 3900
acataagata cctattccaa aactgaattc cttttaactc tcatgaaatg acaaatagaa 3960
ttgttagtat atgtgagcac tgagaattac tttattgatg aacactggta ttttctctgg 4020
tgtaggaagc tgaaccatct atctccagaa atgtcttcag aaagtaaaga gcaacataac 4080
gtttcaccca gagactcagc tgaaggaaat gacagttatc catctgggat ccatctggaa 4140

cttcaaaggg aatcaagtac tgacttcaag caatttgaga ccaatgatca atgcagacct 4200
tatcatagga tccttattga gcgtcaagag aaatcagata caaacttcaa ggagtttgtt 4260
attaaaaagc tgcagaagaa ttgccagtgc agtccagcca aagccaaaaa tatgatttta 4320
ggtttccttc ctgttttgca gtggctccca aaatacgacc taaagaaaaa cattttaggg 4380
gatgtgatgt caggcttgat tgtgggcata ttattggtgc cccagtccat tgcttattcc 4440
ctgctggctg gccaagaacc tgtctatggt ctgtacacat ctttttttgc cagcatcatt 4500
tattttctct tgggtacctc ccgtcacatc tctgtgggca tttttggagt actgtgcctt 4560
atgattggtg agacagttga ccgagaacta cagaaagctg gctatgacaa tgcccatagt 4620
gctccttcct taggaatggt ttcaaatggg agcacattat taaatcatac atcagacagg 4680
atatgtgaca aaagttgcta tgcaattatg gttggcagca ctgtaacctt tatagctgga 4740
gtttatcagg taagcagcaa tgaaacaatt ggttatttct agaaaagtaa tctagtacat 4800
gaaatctcat atctctaagg gatctgagga atcacaataa ttaaaggtat catttattga 4860
gagttcagga tatatgaagg gtagaggcaa aattcaaacc ctaacctgac tccacaggta 4920
atataaggct ggttcactgg acctccacca cccagtacaa ctccttaatt ttacatgtca 4980
gaaaatcttg gctttgcttg agattatttg tggctggtta ttggcagagt cagcattagc 5040
agttaggcaa gtgggtaaca gaatggagtt gagagtgcag gagtttctca ctttttttt 5100
ttttctggag acagggtctc actctgtcac gctggagtgc agtggcacta tcttagttca 5160
ctgcaacgtc cgcctccctg gctcaagcag tcctcctacc tcaacctcct gagtagctag 5220
gactacaggc acatgctacc acacctggct aattttattt tattttattt tattttattt 5280
tttatttta ttttttgtag agacagggtt ttgccacgtt gcccaggctg gtttcaaact 5340
cctgagctca agcaatcctc ccgtcttggc ctcccaaagt gctgggatta tagccatgag 5400
ccaccacacc cagcctcaaa ttctaaatgt ctcttacctt ccattaaaat tgctgatcta 5460
ttgagcaact cttactaaag gtagtggttg tcttggattg ttggggaggg agggaaaaag 5520
ttggggacca cagtttcata ttatcagcca ggagaaagga taagaaatca aattcttgag 5580
tctcccatag aatccactaa tctgtcatta tcatcatgcc cctggctttt ggcatccagg 5640
agtcagtgcc aggattaaac cttctctaat gcaggcattt caaaccaaca agggaagggg 5700
aagagtagct cactttagtt ggtgctcaga tgagtgggga gggagagtga agatggtgtg 5760
aagatgagct gtctactcat atataatggt aaataataag tctacttact tatttattat 5820
ttattcattt atttataaag agacagggtc tctctatgac caaactcctg ggctcaagtg 5880
atcctcctaa tattgcctcc ccaaatgctg ggattacagg catgagccat cacgcccaac 5940
caacttttgc ctttttgtta gtatgtccca ccaagaagga agaaggcata acaattctga 6000
aaacttatta gacagaggaa aatataaaga agtaaaaatg cagaattttt attaatatgg 6060
gagacagtgt ggcataagta catatatact gcatgagaat ggtttcttag tatgaggtta 6120
aagataagtc tacaataatt tttaaagtgt gattctactt tgatgtaaat ctaatttttt 6180
gttttaccaa ttaaaacttc acttgtacac ttgctcttag ccaagaggct gagaagccgt 6240
aagacttcac ttttacagta gtgatttgta atttaaggaa aatacttggt ttcttaacta 6300
gaataatttt ttccaatttg aagttttctt gtggatcctt gagaatgttt ttcttttaaa 6360
agaggtctgt tctttgtgat gggaagaatg aaaaaaaaaa gaggtatgaa ccttattcaa 6420
gtttaagaaa cgtatgaaaa gaaagaaatc caaagttcct gtctcacctg ggttaataag 6480
taacagtgtg accttgggca agttgcttag ccctttaaac ataattttca tctttgtaaa 6540
atgagaagat tgatatatga ttgtgtttat tctagctctg acattctgtg atgctctgat 6600
gatatgtctc catgcaagaa atgtcaggat aatataaaat ttagaagttc ttttccattt 6660
atatttaaca cttctatatc cttccttcca ggtagcgatg ggcttctttc aagtgggttt 6720
tgtttctgtc tacctctcag atgccttgct gagtggattt gtcactggtg cctccttcac 6780
tattcttaca tctcaggcca agtatcttct tgggctcaac cttcctcgga ctaatggtgt 6840
gggctcactc atcactacct ggatacatgt cttcagaaac atccataaga ccaatctctg 6900
tgatcttatc accagccttt tgtgcctttt ggttcttttg ccaaccaaag aactcaatga 6960
acacttcaaa tccaagctta aggcaccgat tcctattgaa cttgttgttg ttgtagcagc 7020

cacattagcc tctcattwtg gaaaactaca tgaaaattat aattctagta ttgctggaca 7080
tattcccact gggtttatgc cacccaaagt accagaatgg aacctaattc ctagtgtggc 7140
tgtagatgca atagctattt ccatcattgg ttttgctatc actgtatcac tttctgagat 7200
gtttgccaag aaacatggtt acacagtcaa agcaaaccag gaaatgtatg ccattggctt 7260
ttgtaatatc atcccttcct tcttccactg ttttactact agtgcagctc ttgcaaagac 7320
attggttaaa gaatcaacag gctgccatac tcagctttct ggtgtggtaa cagccctggt 7380
tcttttgttg gtcctcctag taatagctcc tttgttctat tcccttcaaa aaagtgtcct 7440
tggtgtgatc acaattgtaa atctacgggg agcccttcgt aaatttaggg atcttcccaa 7500
aatgtggagt attagtagaa tggatacagt tatctggttt gttactatgc tgtcctctgc 7560
actgctaagt actgaaatag gcctacttgt tggggtttgt ttttctatat tttgtgtcat 7620
cctccgcact cagaagccaa agagttcact gcttggcttg gtggaagagt ctgaggtctt 7680
tgaatctgtg tctgcttaca agaaccttca gaytaagcca ggcatcaaga ttttccgctt 7740
tgtagcccct ctctactaca taaacaaaga atgctttaaa tctgctttat acaaacaaac 7800
tgtcaaccca atcttaataa aggtggcttg gaagaaggca gcaaagagaa agatcaaaga 7860
aaaagtagtg actcttggtg gaatccagga tgaaatgtca gtgcaacttt cccatgatcc 7920
cttggagctg catactatag tgattgactg cagtgcaatt caatttttag atacagcagg 7980
gatccacaca ctgaaagaag ttcgcagaga ttatgaagcc attggaatcc aggttctgct 8040
ggctcagtgc aatcccwctg tgagggattc cctaaccaac ggagaatatt gcaaaaagga 8100
agaagaaaac cttctcttct atagtgtgta tgaagcgatg gcttttgcag aagtatctaa 8160
aaatcagaaa ggagtatgtg ttcccaatgg tctgagtctt agtagtgatt aattgagaag 8220
gtagatagaa gaatgtctag ccaataggtt aaaatttcaa gtgtccaaca tttcccagtt 8280
ccacagtggg aaattttgca cacttgaaat tttaaccaag tggctagata ttattcctcc 8340
tttgaagcta atggcatttg tatatacaca ctgcagcaga gcttgtagct ggacagagtc 8400
aaaaagaaga aaatacggtt tcaggctttc ttgcagatat gaagtattct tggaatgcaa 8460
taagtatgta ttgaactgta ctgtaaagta gctccaaaac ttaattactc tcctgtttta 8520
ggggttatac atttggactg tgcattctcc aagagatgaa gcggtgaagt tgggatttac 8580
attggaagtg ctgtagactt ctttatgtgg ctcagtggag agagggaaag aatgttgcac 8640
ctgctctagt accataggtc aagaggcttc tggatcacaa agtcataact agacaggttt 8700
gttcttgtag ttttctatcc ccagtctttg ctccccagat ggcagtagtt tttagtagga 8760
aagtgccatt cctgtcctta aggcacagtc tcatcagaag tctaatacct gggcaggttt 8820
ataacatcct gagagccagc ctgacattag acagaatacc ctttgtaata cattggaaat 8880
ttttactcat gccttttttgt ttaggataaa taggtaagca caaagagctc ttcaaaatca 8940
gaaaaaacaa taggagtcct tccttgtctt ttctgtgatc tctgtccttg tttctgagac 9000
tttctctacc attaagctct attttagctt tcagttattc tagtttgttt cccatggaat 9060
ctgtcctaaa ctggtgtttt tgtcagtgac agtcttgcca gtcagcaatt tctaacagca 9120
ttttaaatga gtttgatgta cagtaaatat tgatgacaat gacagctttt aactcttcaa 9180
gtcacctaaa gctattatgc aggaggattt agaagtcaca ttcataaaac ccaagggcta 9240
tgggtgtatt attcatgata gctggcccac aggtcatgaa ttgaggagga atttgctttc 9300
aaaaagcaag aatgtccaac actgaaagtt tatagtttta tatttggacc ttgaaaggta 9360
agaaaaaacc aggttctcca aagttaggaa tagggaacta atttatgaaa cagccatctt 9420
aaaaaaaaaa aaagtaaact gcaaaagtac aaaatcattt ttcaatctgt tcccagtttc 9480
taaacaattt taaatattta tgagaagcaa accctatgtg tagggcatct gttggagtgg 9540
gatgctttta gacatatatt aagtatgtac atgtttaata tgtatattta aaatgcatat 9600
atattttatt atatctatat tatcctatat agatatatgt aacttagctt tattgttagc 9660
tccataagct gccagtgttg cttttctgtt ggtagagctc tcccatttgg tgacatggaa 9720
aataccttte cattatcaca acaaagcagt tgctcagtag aaagtctaga tttctgtctt 9780
ataggtgatt tctgtcttat aggtgattat aatcaagtgt aggcttcctg aattttgaca 9840
tccttttaga acttgggtct ggaattccag aaatgttaat tgctgcttgt atttgttctt 9900

gtttgttttt tagccagtat ttgccctttc tatccagcct tatgaataat agcagtaaaa 9960
tcacagtatc ttggtcagtc tttatttttt tccttttttc ttttttaaga gacagtcatc 10020
caggccagag tgcagtttga tgatagctta ctgaagcttc ccactcctgg gctcaagtta 10080
tccttccatt ttggcctcct gagtagctag accataggta tgcatcacca caccctgcta 10140
atttttaaa ttttttttcta gagagagggt ctcactgtgt tgcccaggct ggtctcaaac 10200
tccaggctca agcaatcctt cagcctcagc ctcccagagt gttgggatta caggcgtgag 10260
ccactgcact tggccaagtt atttatttttt aatctctctt gcccttctcc caaggcaggc 10320
ttaagttgag actattatag gtgtctaata acctgtgaca gagtaatgag tacatgctta 10380
agatgttata attagccaac accaacacag caaaaaatat aattccagcc aaagattctg 10440
gaaaatccct cagaaggagg gataacagga tttgaccttt accagcgatt tctgtccata 10500
tgtggatgta aacagttctg gaacgttatg catgcagtta gcgaatcctt gaattatgtt 10560
ctggtttgta cttgtcccat ccatccaaac aagagattct gcttttggta gccatctgta 10620
gaaacattta agatgtcact agaatttaca tttcatcctc tctacttggg ttgaggttgc 10680
ctatacttgc atattgttaa aatgttttgg ttgctgatat tcagaggaat gaaacctgga 10740
accaaagcct aatttgccga taaaaaaact gttttcggcc aggtgcagtg gctcatgcct 10800
gtaatcccag cacgttggga ggccgaggcg ggtggatcac ctgaagtcag gagttcgaga 10860
ccatcctggc taacactgtg aaaccccgtc tctactaaaa atacaaaaaa ttagcggggc 10920
atggtggcac gcgcctgtag tcccagctac tcaggaggct gaggcaggag aattacttga 10980
acccgggagg cggaggttgc agtaagcaga gattgtgcca ctggactcca gcctgggtga 11040
cagagcgaga ctccgtctca aacaaacaaa caaaaaactg ttttcatttg ctctcttgac 11100
caaaggatag gactttagtt ctttaagcat tattttaaac actatattga tacaaaaata 11160
tcttgcttac tctaaacttt agagtctaaa tgaagctttt tctcagtaca agattctgag 11220
tatcataaaa tggttattta attgaaacgt agtgtggtat actcttgatg gttagaactc 11280
ttacagcctt atttattttt aagtttgtta cagccaaagg gttggagtgt gccagtgcac 11340
aggtagacta aggaaaacat tatagaggag tgaagagaac agaccattga aaagactatt 11400
atctgaccag cggaggcaga aaagagagga acccagttga ataggatcca atccctggtt 11460
agcctctaca caataatagg gagacaagga ttaggagcca tacctcccag agcaaggtat 11520
ctttctagag caaatttctc tttctagaag gggagggtca cagggtcaca gattcaccaa 11580
agctgaaagg gctgaggagc tcatggtagc ctgggttgac ctactctgga gcacggtgtc 11640
ttccttctaa actgagtgac tgtagtacta tctgtgcctc tgatggtaat aaaactgaca 11700
agatgtctaa ttttttttta agtaggacca aaggaaaaca agatttagat agtctgactt 11760
tgcttttgaa caacagacat tgcaagtcaa aattgttgtc aaatttacat atggtaaatg 11820
atgaacttta aaaatgtgtc caggtgttag atgagttcat tagactcttt taatgctaat 11880
ggctagtacg tttaaacaaa acagcagttc tctgctgcaa tattcccatt gaccacttaa 11940
atgaccataa gtggtcattt aagaacatgt tagggttagc cctgatctga atataaaagt 12000
gagaaaaggg ctacagtgca tttcttggta acttaaactg agtcttgaag ttataatgat 12060
ccattcgagt tctgtgatcc ttattgttct taattgtgtt tctctacgta ttgttacaga 12120
tgagccatac gtttctttgt atcaatgtag acatgacttc agatacctct gaggacctac 12180
ccagcagtct aggaccctgg gccaagtgct gg 12212

<210> 2
<211> 2220
<212> DNA
<213> Homo sapiens

<400> 2
atgtcttcag aaagtaaaga gcaacataac gtttcaccca gagactcagc tgaaggaaat 60

gacagttatc catctgggat ccatctggaa cttcaaaggg aatcaagtac tgacttcaag 120
caatttgaga ccaatgatca atgcagacct tatcatagga tccttattga gcgtcaagag 180
aaatcagata caaacttcaa ggagtttgtt attaaaaagc tgcagaagaa ttgccagtgc 240
agtccagcca aagccaaaaa tatgatttta ggtttccttc ctgttttgca gtggctccca 300
aaatacgacc taaagaaaaa cattttaggg gatgtgatgt caggcttgat tgtgggcata 360
ttattggtgc cccagtccat tgcttattcc ctgctggctg gccaagaacc tgtctatggt 420
ctgtacacat ctttttttgc cagcatcatt tattttctct tgggtacctc ccgtcacatc 480
tctgtgggca tttttggagt actgtgcctt atgattggtg agacagttga ccgagaacta 540
cagaaagctg gctatgacaa tgcccatagt gctccttcct taggaatggt ttcaaatggg 600
agcacattat taaatcatac atcagacagg atatgtgaca aaagttgcta tgcaattatg 660
gttggcagca ctgtaacctt tatagctgga gtttatcagg tagcgatggg cttctttcaa 720
gtgggttttg tttctgtcta cctctcagat gccttgctga gtggatttgt cactggtgcc 780
tccttcacta ttcttacatc tcaggccaag tatcttcttg ggctcaacct tcctcggact 840
aatggtgtgg gctcactcat cactacctgg atacatgtct tcagaaacat ccataagacc 900
aatctctgtg atcttatcac cagccttttg tgcctttttgg ttcttttgcc aaccaaagaa 960
ctcaatgaac acttcaaatc caagcttaag gcaccgattc ctattgaact tgttgttgtt 1020
gtagcagcca cattagcctc tcattttgga aaactacatg aaaattataa ttctagtatt 1080
gctggacata ttcccactgg gtttatgcca cccaaagtac cagaatggaa cctaattcct 1140
agtgtggctg tagatgcaat agctatttcc atcattggtt ttgctatcac tgtatcactt 1200
tctgagatgt ttgccaagaa acatggttac acagtcaaag caaaccagga aatgtatgcc 1260
attggctttt gtaatatcat cccttccttc ttccactgtt ttactactag tgcagctctt 1320
gcaaagacat tggttaaaga atcaacaggc tgccatactc agctttctgg tgtggtaaca 1380
gccctggttc ttttgttggt cctcctagta atagctcctt tgttctattc ccttcaaaaa 1440
agtgtccttg gtgtgatcac aattgtaaat ctacggggag cccttcgtaa atttagggat 1500
cttcccaaaa tgtggagtat tagtagaatg gatacagtta tctggtttgt tactatgctg 1560
tcctctgcac tgctaagtac tgaaataggc ctacttgttg gggtttgttt ttctatattt 1620
tgtgtcatcc tccgcactca gaagccaaag agttcactgc ttggcttggt ggaagagtct 1680
gaggtctttg aatctgtgtc tgcttacaag aaccttcaga ctaagccagg catcaagatt 1740
ttccgctttg tagccccttt ctactacata aacaaagaat gctttaaatc tgctttatac 1800
aaacaaactg tcaacccaat cttaataaag gtggcttgga agaaggcagc aaagagaaag 1860
atcaaagaaa aagtagtgac tcttggtgga atccaggatg aaatgtcagt gcaactttcc 1920
catgatccct tggagctgca tactatagtg attgactgca gtgcaattca atttttagat 1980
acagcaggga tccacacact gaaagaagtt cgcagagatt atgaagccat tggaatccag 2040
gttctgctgg ctcagtgcaa tcccactgtg agggattccc taaccaacgg agaatattgc 2100
aaaaaggaag aagaaaacct tctcttctat agtgtgtatg aagcgatggc ttttgcagaa 2160
gtatctaaaa atcagaaagg agtatgtgtt cccaatggtc tgagtcttag tagtgattaa 2220

<210> 3
<211> 739
<212> PRT
<213> Homo sapiens

<400> 3
Met Ser Ser Glu Ser Lys Glu Gln His Asn Val Ser Pro Arg Asp Ser
1 5 10 15

Ala Glu Gly Asn Asp Ser Tyr Pro Ser Gly Ile His Leu Glu Leu Gln

20 25 30

Arg Glu Ser Ser Thr Asp Phe Lys Gln Phe Glu Thr Asn Asp Gln Cys
35 40 45

Arg Pro Tyr His Arg Ile Leu Ile Glu Arg Gln Glu Lys Ser Asp Thr
50 55 60

Asn Phe Lys Glu Phe Val Ile Lys Lys Leu Gln Lys Asn Cys Gln Cys
65 70 75 80

Ser Pro Ala Lys Ala Lys Asn Met Ile Leu Gly Phe Leu Pro Val Leu
85 90 95

Gln Trp Leu Pro Lys Tyr Asp Leu Lys Lys Asn Ile Leu Gly Asp Val
100 105 110

Met Ser Gly Leu Ile Val Gly Ile Leu Leu Val Pro Gln Ser Ile Ala
115 120 125

Tyr Ser Leu Leu Ala Gly Gln Glu Pro Val Tyr Gly Leu Tyr Thr Ser
130 135 140

Phe Phe Ala Ser Ile Ile Tyr Phe Leu Leu Gly Thr Ser Arg His Ile
145 150 155 160

Ser Val Gly Ile Phe Gly Val Leu Cys Leu Met Ile Gly Glu Thr Val
165 170 175

Asp Arg Glu Leu Gln Lys Ala Gly Tyr Asp Asn Ala His Ser Ala Pro
180 185 190

Ser Leu Gly Met Val Ser Asn Gly Ser Thr Leu Leu Asn His Thr Ser
195 200 205

Asp Arg Ile Cys Asp Lys Ser Cys Tyr Ala Ile Met Val Gly Ser Thr
210 215 220

Val Thr Phe Ile Ala Gly Val Tyr Gln Val Ala Met Gly Phe Phe Gln
225 230 235 240

Val Gly Phe Val Ser Val Tyr Leu Ser Asp Ala Leu Leu Ser Gly Phe
245 250 255

Val Thr Gly Ala Ser Phe Thr Ile Leu Thr Ser Gln Ala Lys Tyr Leu
260 265 270

Leu Gly Leu Asn Leu Pro Arg Thr Asn Gly Val Gly Ser Leu Ile Thr

275 280 285

Thr Trp Ile His Val Phe Arg Asn Ile His Lys Thr Asn Leu Cys Asp
290 295 300

Leu Ile Thr Ser Leu Leu Cys Leu Leu Val Leu Leu Pro Thr Lys Glu
305 310 315 320

Leu Asn Glu His Phe Lys Ser Lys Leu Lys Ala Pro Ile Pro Ile Glu
325 330 335

Leu Val Val Val Val Ala Ala Thr Leu Ala Ser His Phe Gly Lys Leu
340 345 350

His Glu Asn Tyr Asn Ser Ser Ile Ala Gly His Ile Pro Thr Gly Phe
355 360 365

Met Pro Pro Lys Val Pro Glu Trp Asn Leu Ile Pro Ser Val Ala Val
370 375 380

Asp Ala Ile Ala Ile Ser Ile Ile Gly Phe Ala Ile Thr Val Ser Leu
385 390 395 400

Ser Glu Met Phe Ala Lys Lys His Gly Tyr Thr Val Lys Ala Asn Gln
405 410 415

Glu Met Tyr Ala Ile Gly Phe Cys Asn Ile Ile Pro Ser Phe Phe His
420 425 430

Cys Phe Thr Thr Ser Ala Ala Leu Ala Lys Thr Leu Val Lys Glu Ser
435 440 445

Thr Gly Cys His Thr Gln Leu Ser Gly Val Val Thr Ala Leu Val Leu
450 455 460

Leu Leu Val Leu Leu Val Ile Ala Pro Leu Phe Tyr Ser Leu Gln Lys
465 470 475 480

Ser Val Leu Gly Val Ile Thr Ile Val Asn Leu Arg Gly Ala Leu Arg
485 490 495

Lys Phe Arg Asp Leu Pro Lys Met Trp Ser Ile Ser Arg Met Asp Thr
500 505 510

Val Ile Trp Phe Val Thr Met Leu Ser Ser Ala Leu Leu Ser Thr Glu
515 520 525

Ile Gly Leu Leu Val Gly Val Cys Phe Ser Ile Phe Cys Val Ile Leu

530 535 540

Arg Thr Gln Lys Pro Lys Ser Ser Leu Leu Gly Leu Val Glu Glu Ser
545 550 555 560

Glu Val Phe Glu Ser Val Ser Ala Tyr Lys Asn Leu Gln Thr Lys Pro
565 570 575

Gly Ile Lys Ile Phe Arg Phe Val Ala Pro Leu Tyr Tyr Ile Asn Lys
580 585 590

Glu Cys Phe Lys Ser Ala Leu Tyr Lys Gln Thr Val Asn Pro Ile Leu
595 600 605

Ile Lys Val Ala Trp Lys Lys Ala Ala Lys Arg Lys Ile Lys Glu Lys
610 615 620

Val Val Thr Leu Gly Gly Ile Gln Asp Glu Met Ser Val Gln Leu Ser
625 630 635 640

His Asp Pro Leu Glu Leu His Thr Ile Val Ile Asp Cys Ser Ala Ile
645 650 655

Gln Phe Leu Asp Thr Ala Gly Ile His Thr Leu Lys Glu Val Arg Arg
660 665 670

Asp Tyr Glu Ala Ile Gly Ile Gln Val Leu Leu Ala Gln Cys Asn Pro
675 680 685

Thr Val Arg Asp Ser Leu Thr Asn Gly Glu Tyr Cys Lys Lys Glu Glu
690 695 700

Glu Asn Leu Leu Phe Tyr Ser Val Tyr Glu Ala Met Ala Phe Ala Glu
705 710 715 720

Val Ser Lys Asn Gln Lys Gly Val Cys Val Pro Asn Gly Leu Ser Leu
725 730 735

Ser Ser Asp

<210> 4
<211> 15
<212> DNA
<213> Homo sapiens

<400> 4

aagtcctrta cccag 15

<210> 5
<211> 15
<212> DNA
<213> Homo sapiens

<400> 5
ttaaggaraa gggac 15

<210> 6
<211> 15
<212> DNA
<213> Homo sapiens

<400> 6
tctcattwtg gaaaa 15

<210> 7
<211> 15
<212> DNA
<213> Homo sapiens

<400> 7
caatcccwct gtgag 15

<210> 8
<211> 15
<212> DNA
<213> Homo sapiens

<400> 8
cttgggaagt cctrt 15

<210> 9
<211> 15
<212> DNA
<213> Homo sapiens

<400> 9
aactggctgg gtaya 15

<210> 10
<211> 15
<212> DNA
<213> Homo sapiens

<400> 10
gctcaattaa ggara 15

<210> 11
<211> 15
<212> DNA
<213> Homo sapiens

<400> 11
tcttatgtcc cttyt 15

<210> 12
<211> 15
<212> DNA
<213> Homo sapiens

<400> 12
ttagcctctc attwt 15

<210> 13
<211> 15
<212> DNA
<213> Homo sapiens

<400> 13
atgtagtttt ccawa 15

<210> 14
<211> 15
<212> DNA
<213> Homo sapiens

<400> 14
tcagtgcaat cccwc 15

<210> 15
<211> 15
<212> DNA

<213> Homo sapiens

<400> 15
gaatccctca cagwg 15

<210> 16
<211> 10
<212> DNA
<213> Homo sapiens

<400> 16
gggaagtcct 10

<210> 17
<211> 10
<212> DNA
<213> Homo sapiens

<400> 17
tggctgggta 10

<210> 18
<211> 10
<212> DNA
<213> Homo sapiens

<400> 18
caattaagga 10

<210> 19
<211> 10
<212> DNA
<213> Homo sapiens

<400> 19
tatgtccctt 10

<210> 20
<211> 10
<212> DNA
<213> Homo sapiens

<400> 20

gcctctcatt 10

<210> 21
<211> 10
<212> DNA
<213> Homo sapiens

<400> 21
tagttttcca 10

<210> 22
<211> 10
<212> DNA
<213> Homo sapiens

<400> 22
gtgcaatccc 10

<210> 23
<211> 10
<212> DNA
<213> Homo sapiens

<400> 23
tccctcacag 10

<210> 24
<211> 12212
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (3798)
<223> PS1: Polymorphic base G or A

<220>
<221> allele
<222> (3895)
<223> PS2: Polymorphic base A or G

<220>
<221> allele
<222> (7038)

<223> PS3: Polymorphic base T or A

<220>
<221> allele
<222> (7713)
<223> PS4: Polymorphic base C or T

<220>
<221> allele
<222> (8057)
<223> PS5: Polymorphic base A or T

<400> 24
gtgtttatct gagaaagttt ttatttcttt ttcatatttt attttattta agacagagtc 60
tctccctgtc acccaggcta aagtgcagtg gtgtgatcat ggcttattgc agcctcaacc 120
tcttaggctc aagcagtctc cctacctcag cctcccaagt agctgggacc ataggcacac 180
accaccatgc ctggcctctt atcttttaaa atatttttc taggtaaaga attctatgtt 240
aaaagttata tttctggggt cagatgtggt ggctcacacc tgtaatccta gcactttggg 300
aggccagggt gggaggatca cttgaggcca agagtttggg aaccagcctg gggagcttag 360
tagcaagacc ttgtctctac aaaaaaaatt aaaaaatgag ccaggcttgg tcgtgcacac 420
ctgtaatcct cgctactcag gaggctgagg tgggagaatc acttgagccc aggagttcga 480
ggccatggtg agccataatt gcatgtcact gcacttcagc ctgggcaaca gggcaagacc 540
ctatttctat aaaagatata tatttctttg atacaaagat accactccat catgttctgg 600
cctccatgat ttccaataag aagtctgcta taattcttat ttttgtttct gtataagtaa 660
tatatgtttt ttctttgtct tcaggatttt ctctttatct tttgttttca gcagcttgaa 720
tatcatatgt gtaggtgagg gtttttgttt tctttttgtt tttatattta ctcttcttgg 780
aattctctaa gattcttgga gatagattat atattatata tatatgtatt tttttttga 840
gacagagttt tgctttgtca ccatgctgga gtgcagtggc gcgatcttgg ctcactgcag 900
cctccgcctc ccaggttcaa gcaattctcc tgcctcagcc tcctgagtag ctgggactac 960
aggagcgcac caccatgccc agctaatttt tgtatttttt tagaagagac ggggtttcac 1020
catgttggcc aggatggtct tgatctcttg accttgtgat ctgcctgcct tggcctccgg 1080
aagtgttgga attacaggtg tgagccactg cacccggcct ttttttttt ttttttttt 1140
gagaaggagt tttcctcttg ctgtccagac aggagtgcaa tggcgtgatc tcggctcact 1200
gcaacctcca gttcctgggt tccagcgatt ctcctgcctc ggcctcgcga gtagctggga 1260
ttacaggcgt gtgccaccat gcctgactaa ttttatattt ttagtagaga ccctgtttca 1320
ccatgttggt catgctggtc ttgaactcct ggcctcaggt tacttggatc tatattttga 1380
tgtctcttat taattttgga aatttcttgg ccattatatc ttgaaatatt tcttctgccc 1440
attctctctc ttcttttggt attctaatta tgcatatgtt tgatactttg atgtgtcccg 1500
cagctcttgg atgttatgtt acattgttga tattggggtt tttttgtttt gtttcgttat 1560
ttaaactttt ctgcgtttta gtttagctaa tttctgttta tctgtctttt tgagttcatc 1620
gattctttct ttggctttcg ggtctagtaa tgatcctatc gaaggcattc atcatctctg 1680
ttactgtgtt tttcactttt agcattcctt atttgatttt ttcttacagt ttccatttct 1740
ttgttttact ttctatgtga atttttatct atattcttcc cagatagcag ccatactggg 1800
caagatttgg attcccttag ttttttctta tgtctctttt gcatatctta ggtttggaaa 1860
aacctttca cccagaaatc accaagagct ctccaaaata ggacgttggt gcactcaagc 1920
tagctctaga aagcaggaaa tgacaggaga gtcacagact ggttgcattg tagagcaaag 1980
agctgctggg tgtatccatc aacctgattg attttctcag ggtcagtggt cacctaagtc 2040
tattctttct ttccttatct acttcagaag accatcaaaa tatttactca tgttaaacct 2100

aaacaaaatg aatatgggtt acttttacat atcttgtgtt cttgaataag tcataggtcc 2160
tggattttgg cagtttggta acgtaaggtt tatagaactt cacaagggcc cattttctg 2220
tcagttccaa aacaagaaac ttcctgagtc cttcatctaa aatatattgt ctcggctggg 2280
tgcagtggct cgtgcctgtt atcccagcac tttgggaggc tgaggcggat ggatcacctg 2340
aggttaggac ttcgagacca gcctgaccaa tatggtgaaa ccctgtctct actaaaaatg 2400
caaaaattag ccaggcgtgg tggcatgcac ctgtaatccc agctactcag gaggctgaga 2460
taagagaatt gcttgaaccc aggaggtgga ggttgtagtg agctgagatt gcaccactgc 2520
actccagcct gggcgacaga gtgaaactct gtctcaaaaa aaaaaaaaaa aatgtctcta 2580
tctggccaca gtcacaaatg tttgttcatt tgttcattca ttcattcaaa tgttttgtaa 2640
gcctgctatc tcagcgttac tacattccat tcagattaca ctgatgaaca agatgtcttt 2700
cctccaggag ctagagagat tcctacttca ctaatacaag agtgtggtta gtactctaat 2760
ggaggtgcaa catgctatgg gacacagagg gtgtagtatt tcatttgggc taggggagat 2820
tggttagtgc tttctggaaa aggtagcatt gtaactgggt tttaaaaaat tattaggatc 2880
ttgacaggca aagaggtgga tggccattcg aagctaagta aacagcttat gtaaaggcac 2940
taattcatga agcatttggt aaacaattta tgttctattc ctttgagagc ctggttcatt 3000
ttcttctctt actccggtta ttagacttac tatttgttgt tgtcctttct ctttttctgg 3060
ctatttttac ctcctttgtt ttcctatagt tcctcatggt agatcttatg gcattagttt 3120
tatagtctag gacacagaga tgaaggatca cctgtattgc ctccaagtgg aagtgcaggg 3180
caacattatt tctctattta acctgtgttt cagtgtgtgt acttagaata gtaaagtgaa 3240
tcttgcatga atgtaggccc tgcccacagg gcagatgact ccatactaga acatagtgga 3300
atagacaaaa accttctaca gcatgtatga gacacttggc ccatcgaccc tcttcatgcc 3360
ctttacattc agcaccctca tattgacttc tctctcctct ttcctaccaa gcaagggagt 3420
actgttcaaa gacgcaaatg catttgccc tagtttcttt ttattgctaa aaacatttat 3480
ctttacccta caacctactt ttctatttat tttcaacatt tagcaggttg tttaaaaagg 3540
gaccaaaaaa taaaacagga ccatcttcct tgtttcaggg actggtaggc aggcattaag 3600
gttaaggtag gggttaagac cagatcctat tttgcagtct gcctgggagg tgaaaaacct 3660
gggaagaaga ccgctggtag catatgtatg gaaaggagac aggctgccct tacatctttt 3720
caggaggaaa aactgccagg gggagccagg catatatgga gaagaatcct taatggttta 3780
tactcttggg aagtcctrta cccagccagt tatttgcttt gacttggctg tttaaggtct 3840
ggttctggtc ttttttttte ccccctaacca agacaaatga ggctcaatta aggaraaggg 3900
acataagata cctattccaa aactgaattc cttttaactc tcatgaaatg acaaatagaa 3960
ttgttagtat atgtgagcac tgagaattac tttattgatg aacactggta ttttctctgg 4020
tgtaggaagc tgaaccatct atctccagaa atgtcttcag aaagtaaaga gcaacataac 4080
gtttcaccca gagactcagc tgaaggaaat gacagttatc catctgggat ccatctggaa 4140
cttcaaaggg aatcaagtac tgacttcaag caatttgaga ccaatgatca atgcagacct 4200
tatcatagga tccttattga gcgtcaagag aaatcagata caaacttcaa ggagtttgtt 4260
attaaaaagc tgcagaagaa ttgccagtgc agtccagcca aagccaaaaa tatgatttta 4320
ggtttccttc ctgttttgca gtggctccca aaatacgacc taaagaaaaa cattttaggg 4380
gatgtgatgt caggcttgat tgtgggcata ttattggtgc cccagtccat tgcttattcc 4440
ctgctggctg gccaagaacc tgtctatggt ctgtacacat ctttttttgc cagcatcatt 4500
tattttctct tgggtacctc ccgtcacatc tctgtgggca tttttggagt actgtgcctt 4560
atgattggtg agacagttga ccgagaacta cagaaagctg gctatgacaa tgcccatagt 4620
gctccttcct taggaatggt ttcaaatggg agcacattat taaatcatac atcagacagg 4680
atatgtgaca aaagttgcta tgcaattatg gttggcagca ctgtaacctt tatagctgga 4740
gtttatcagg taagcagcaa tgaaacaatt ggttatttct agaaaagtaa tctagtacat 4800
gaaatctcat atctctaagg gatctgagga atcacaataa ttaaaggtat catttattga 4860
gagttcagga tatatgaagg gtagaggcaa aattcaaacc ctaacctgac tccacaggta 4920
atataaggct ggttcactgg acctccacca cccagtacaa ctccttaatt ttacatgtca 4980

gaaaatcttg gctttgcttg agattatttg tggctggtta ttggcagagt cagcattagc 5040
agttaggcaa gtgggtaaca gaatggagtt gagagtgcag gagtttctca ctttttttt 5100
ttttctggag acagggtctc actctgtcac gctggagtgc agtggcacta tcttagttca 5160
ctgcaacgtc cgcctccctg gctcaagcag tcctcctacc tcaacctcct gagtagctag 5220
gactacaggc acatgctacc acacctggct aattttattt tattttattt tattttattt 5280
tttatttta ttttttgtag agacagggtt ttgccacgtt gcccaggctg gtttcaaact 5340
cctgagctca agcaatcctc ccgtcttggc ctcccaaagt gctgggatta tagccatgag 5400
ccaccacacc cagcctcaaa ttctaaatgt ctcttacctt ccattaaaat tgctgatcta 5460
ttgagcaact cttactaaag gtagtggttg tcttggattg ttggggaggg agggaaaaag 5520
ttggggacca cagtttcata ttatcagcca ggagaaagga taagaaatca aattcttgag 5580
tctcccatag aatccactaa tctgtcatta tcatcatgcc cctggctttt ggcatccagg 5640
agtcagtgcc aggattaaac cttctctaat gcaggcattt caaaccaaca agggaagggg 5700
aagagtagct cactttagtt ggtgctcaga tgagtgggga gggagagtga agatggtgtg 5760
aagatgagct gtctactcat atataatggt aaataataag tctacttact tatttattat 5820
ttattcattt atttataaag agacagggtc tctctatgac caaactcctg ggctcaagtg 5880
atcctcctaa tattgcctcc ccaaatgctg ggattacagg catgagccat cacgcccaac 5940
caacttttgc ctttttgtta gtatgtccca ccaagaagga agaaggcata acaattctga 6000
aaacttatta gacagaggaa aatataaaga agtaaaaatg cagaattttt attaatatgg 6060
gagacagtgt ggcataagta catatatact gcatgagaat ggtttcttag tatgaggtta 6120
aagataagtc tacaataatt tttaaagtgt gattctactt tgatgtaaat ctaatttttt 6180
gttttaccaa ttaaaacttc acttgtacac ttgctcttag ccaagaggct gagaagccgt 6240
aagacttcac ttttacagta gtgatttgta atttaaggaa aatacttggt ttcttaacta 6300
gaataatttt ttccaatttg aagttttctt gtggatcctt gagaatgttt ttcttttaaa 6360
agaggtctgt tctttgtgat gggaagaatg aaaaaaaaaa gaggtatgaa ccttattcaa 6420
gtttaagaaa cgtatgaaaa gaaagaaatc caaagttcct gtctcacctg ggttaataag 6480
taacagtgtg accttgggca agttgcttag ccctttaaac ataattttca tctttgtaaa 6540
atgagaagat tgatatatga ttgtgtttat tctagctctg acattctgtg atgctctgat 6600
gatatgtctc catgcaagaa atgtcaggat aatataaaat ttagaagttc ttttccattt 6660
atatttaaca cttctatatc cttccttcca ggtagcgatg ggcttctttc aagtgggttt 6720
tgtttctgtc tacctctcag atgccttgct gagtggattt gtcactggtg cctccttcac 6780
tattcttaca tctcaggcca agtatcttct tgggctcaac cttcctcgga ctaatggtgt 6840
gggctcactc atcactacct ggatacatgt cttcagaaac atccataaga ccaatctctg 6900
tgatcttatc accagccttt tgtgcctttt ggttcttttg ccaaccaaag aactcaatga 6960
acacttcaaa tccaagctta aggcaccgat tcctattgaa cttgttgttg ttgtagcagc 7020
cacattagcc tctcattwtg gaaaactaca tgaaaattat aattctagta ttgctggaca 7080
tattcccact gggtttatgc cacccaaagt accagaatgg aacctaattc ctagtgtggc 7140
tgtagatgca atagctattt ccatcattgg ttttgctatc actgtatcac tttctgagat 7200
gtttgccaag aaacatggtt acacagtcaa agcaaaccag gaaatgtatg ccattggctt 7260
ttgtaatatc atcccttcct tcttccactg ttttactact agtgcagctc ttgcaaagac 7320
attggttaaa gaatcaacag gctgccatac tcagctttct ggtgtggtaa cagccctggt 7380
tcttttgttg gtcctcctag taatagctcc tttgttctat tcccttcaaa aaagtgtcct 7440
tggtgtgatc acaattgtaa atctacgggg agcccttcgt aaatttaggg atcttcccaa 7500
aatgtggagt attagtagaa tggatacagt tatctggttt gttactatgc tgtcctctgc 7560
actgctaagt actgaaatag gcctacttgt tggggtttgt ttttctatat tttgtgtcat 7620
cctccgcact cagaagccaa agagttcact gcttggcttg gtggaagagt ctgaggtctt 7680
tgaatctgtg tctgcttaca agaaccttca gaytaagcca ggcatcaaga ttttccgctt 7740
tgtagcccct ctctactaca taaacaaaga atgctttaaa tctgctttat acaaacaaac 7800
tgtcaaccca atcttaataa aggtggcttg gaagaaggca gcaaagagaa agatcaaaga 7860

aaaagtagtg actcttggtg gaatccagga tgaaatgtca gtgcaacttt cccatgatcc 7920
cttggagctg catactatag tgattgactg cagtgcaatt caatttttag atacagcagg 7980
gatccacaca ctgaaagaag ttcgcagaga ttatgaagcc attggaatcc aggttctgct 8040
ggctcagtgc aatcccwctg tgagggattc cctaaccaac ggagaatatt gcaaaaagga 8100
agaagaaaac cttctcttct atagtgtgta tgaagcgatg gcttttgcag aagtatctaa 8160
aaatcagaaa ggagtatgtg ttcccaatgg tctgagtctt agtagtgatt aattgagaag 8220
gtagatagaa gaatgtctag ccaataggtt aaaatttcaa gtgtccaaca tttcccagtt 8280
ccacagtggg aaattttgca cacttgaaat tttaaccaag tggctagata ttattcctcc 8340
tttgaagcta atggcatttg tatatacaca ctgcagcaga gcttgtagct ggacagagtc 8400
aaaaagaaga aaatacggtt tcaggctttc ttgcagatat gaagtattct tggaatgcaa 8460
taagtatgta ttgaactgta ctgtaaagta gctccaaaac ttaattactc tcctgtttta 8520
ggggttatac atttggactg tgcattctcc aagagatgaa gcggtgaagt tgggatttac 8580
attggaagtg ctgtagactt ctttatgtgg ctcagtggag agagggaaag aatgttgcac 8640
ctgctctagt accataggtc aagaggcttc tggatcacaa agtcataact agacaggttt 8700
gttcttgtag ttttctatcc ccagtctttg ctccccagat ggcagtagtt tttagtagga 8760
aagtgccatt cctgtcctta aggcacagtc tcatcagaag tctaatacct gggcaggttt 8820
ataacatcct gagagccagc ctgacattag acagaatacc ctttgtaata cattggaaat 8880
ttttactcat gccttttttgt ttaggataaa taggtaagca caaagagctc ttcaaaatca 8940
gaaaaaacaa taggagtcct tccttgtctt ttctgtgatc tctgtccttg tttctgagac 9000
tttctctacc attaagctct attttagctt tcagttattc tagtttgttt cccatggaat 9060
ctgtcctaaa ctggtgtttt tgtcagtgac agtcttgcca gtcagcaatt tctaacagca 9120
ttttaaatga gtttgatgta cagtaaatat tgatgacaat gacagctttt aactcttcaa 9180
gtcacctaaa gctattatgc aggaggattt agaagtcaca ttcataaaac ccaagggcta 9240
tgggtgtatt attcatgata gctggcccac aggtcatgaa ttgaggagga atttgctttc 9300
aaaaagcaag aatgtccaac actgaaagtt tatagtttta tatttggacc ttgaaaggta 9360
agaaaaaacc aggttctcca aagttaggaa tagggaacta atttatgaaa cagccatctt 9420
aaaaaaaaaa aaagtaaact gcaaaagtac aaaatcattt ttcaatctgt tcccagtttc 9480
taaacaattt taaatattta tgagaagcaa accctatgtg tagggcatct gttggagtgg 9540
gatgctttta gacatatatt aagtatgtac atgtttaata tgtatattta aaatgcatat 9600
atattttatt atatctatat tatcctatat agatatatgt aacttagctt tattgttagc 9660
tccataagct gccagtgttg cttttctgtt ggtagagctc tcccatttgg tgacatggaa 9720
aatacctttc cattatcaca acaaagcagt tgctcagtag aaagtctaga tttctgtctt 9780
ataggtgatt tctgtcttat aggtgattat aatcaagtgt aggcttcctg aattttgaca 9840
tccttttaga acttgggtct ggaattccag aaatgttaat tgctgcttgt atttgttctt 9900
gtttgttttt tagccagtat ttgccctttc tatccagcct tatgaataat agcagtaaaa 9960
tcacagtatc ttggtcagtc tttatttttt tccttttttc ttttttaaga gacagtcatc 10020
caggccagag tgcagtttga tgatagctta ctgaagcttc ccactcctgg gctcaagtta 10080
tccttccatt ttggcctcct gagtagctag accataggta tgcatcacca caccctgcta 10140
atttttttaaa ttttttttcta gagagagggt ctcactgtgt tgcccaggct ggtctcaaac 10200
tccaggctca agcaatcctt cagcctcagc ctcccagagt gttgggatta caggcgtgag 10260
ccactgcact tggccaagtt atttatttttt aatctctctt gcccttctcc caaggcaggc 10320
ttaagttgag actattatag gtgtctaata acctgtgaca gagtaatgag tacatgctta 10380
agatgttata attagccaac accaacacag caaaaaatat aattccagcc aaagattctg 10440
gaaaatccct cagaaggagg gataacagga tttgaccttt accagcgatt tctgtccata 10500
tgtggatgta aacagttctg gaacgttatg catgcagtta gcgaatcctt gaattatgtt 10560
ctggtttgta cttgtcccat ccatccaaac aagagattct gcttttggta gccatctgta 10620
gaaacattta agatgtcact agaatttaca tttcatcctc tctacttggg ttgaggttgc 10680
ctatacttgc atattgttaa aatgttttgg ttgctgatat tcagaggaat gaaacctgga 10740

accaaagcct aatttgccga taaaaaaact gttttcggcc aggtgcagtg gctcatgcct 10800
gtaatcccag cacgttggga ggccgaggcg ggtggatcac ctgaagtcag gagttcgaga 10860
ccatcctggc taacactgtg aaaccccgtc tctactaaaa atacaaaaaa ttagcggggc 10920
atggtggcac gcgcctgtag tcccagctac tcaggaggct gaggcaggag aattacttga 10980
acccgggagg cggaggttgc agtaagcaga gattgtgcca ctggactcca gcctgggtga 11040
cagagcgaga ctccgtctca aacaaacaaa caaaaaactg ttttcatttg ctctcttgac 11100
caaaggatag gactttagtt ctttaagcat tattttaaac actatattga tacaaaaata 11160
tcttgcttac tctaaacttt agagtctaaa tgaagctttt tctcagtaca agattctgag 11220
tatcataaaa tggttattta attgaaacgt agtgtggtat actcttgatg gttagaactc 11280
ttacagcctt atttattttt aagtttgtta cagccaaagg gttggagtgt gccagtgcac 11340
aggtagacta aggaaaacat tatagaggag tgaagagaac agaccattga aaagactatt 11400
atctgaccag cggaggcaga aaagagagga acccagttga ataggatcca atccctggtt 11460
agcctctaca caataatagg gagacaagga ttaggagcca tacctcccag agcaaggtat 11520
ctttctagag caaatttctc tttctagaag gggagggtca cagggtcaca gattcaccaa 11580
agctgaaagg gctgaggagc tcatggtagc ctgggttgac ctactctgga gcacggtgtc 11640
ttccttctaa actgagtgac tgtagtacta tctgtgcctc tgatggtaat aaaactgaca 11700
agatgtctaa ttttttttta agtaggacca aaggaaaaca agatttagat agtctgactt 11760
tgcttttgaa caacagacat tgcaagtcaa aattgttgtc aaatttacat atggtaaatg 11820
atgaacttta aaaatgtgtc caggtgttag atgagttcat tagactcttt taatgctaat 11880
ggctagtacg tttaaacaaa acagcagttc tctgctgcaa tattcccatt gaccacttaa 11940
atgaccataa gtggtcattt aagaacatgt tagggttagc cctgatctga atataaaagt 12000
gagaaaaggg ctacagtgca tttcttggta acttaaactg agtcttgaag ttataatgat 12060
ccattcgagt tctgtgatcc ttattgttct taattgtgtt tctctacgta ttgttacaga 12120
tgagccatac gtttctttgt atcaatgtag acatgacttc agatacctct gaggacctac 12180
ccagcagtct aggaccctgg gccaagtgct gg 12212

<210> 25
<211> 540
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (30)
<223> PS1: Polymorphic base G or A

<220>
<221> misc_feature
<222> (61)..(120)
<223> nucleotides represent sequence between PS1 and PS2

<220>
<221> allele
<222> (150)
<223> PS2: Polymorphic base A or G

<220>

<221> misc_feature
<222> (181)..(240)
<223> nucleotides represent sequence between PS2 and PS3

<220>
<221> allele
<222> (270)
<223> PS3: Polymorphic base T or A

<220>
<221> misc_feature
<222> (301)..(360)
<223> nucleotides represent sequence between PS3 and PS4

<220>
<221> allele
<222> (390)
<223> PS4: Polymorphic base C or T

<220>
<221> misc_feature
<222> (421)..(480)
<223> nucleotides represent sequence between PS4 and PS5

<220>
<221> allele
<222> (510)
<223> PS5: Polymorphic base A or T

<400> 25
cttaatggtt tatactcttg ggaagtcctr tacccagcca gttatttgct ttgacttggc 60
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 120
aaccaagaca aatgaggctc aattaaggar aagggacata agatacctat tccaaaactg 180
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 240
tgttgtagca gccacattag cctctcattw tggaaaacta catgaaaatt ataattctag 300
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 360
atctgtgtct gcttacaaga accttcagay taagccaggc atcaagattt tccgctttgt 420
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 480
tccaggttct gctggctcag tgcaatcccw ctgtgaggga ttccctaacc aacggagaat 540

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/20028

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :Please See Extra Sheet.
US CL :Please See Extra Sheet.
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 536/23.1, 23.4; 435/6, 320.1, 325, 455; 514/44

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WEST, MEDLINE, BIOSIS, EMBASE, SCISEARCH, CAPLUS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,830,850 A (GELB et al) 03 November 1998, see entire document.	1-34
Y,P	US 6,143,878 A (KOOPMAN et al) 07 November 2000, see entire document.	1-34
Y	US 5,874,212 A (PROCKOP et al) 23 February 1999, see entire document.	1-34

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
01 OCTOBER 2001

Date of mailing of the international search report
19 NOV 2001

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
PETER PARAD

Telephone No. (703) 308-8340

Form PCT/ISA/210 (second sheet) (July 1998)★

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/20028

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (7):

C07H 21/02, 21/04; C12Q 1/68; C12N 5/00, 5/02, 15/00, 15/09, 15/63, 15/70, 15/74, 15/85, 15/87; A01N 43/04; A61K 31/70

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

536/23.1, 23.4; 435/6, 320.1, 325, 455; 514/44